ON THE COMPARISON OF ONE OR MORE SETS OF PROBABILITY FORECASTS

By Irving I. Gringorten

Geophysics Research Directorate, Air Force Cambridge Research Center

(Original manuscript received 16 August 1957; revised manuscript received 19 December 1957)

ABSTRACT

General expressions for the expected score for accuracy, score for skill and for operational value of the forecasts are developed and discussed. The expressions are then applied to the special case of two-class predictors and predictand, and an example is given to illustate how one set of probability forecasts can meet the operational requirements better than another set, even though both sets of forecasts are equally accurate and equally skillful.

1. Introduction

It has been stated frequently that operational requirements for a weather forecast might be met by a probability statement. For one operation the best working assumption might be "rain," while for another simultaneous operation the better working assumption might be "no rain." But the forecaster can serve *all* operations by stating the probability of rain—or so it is claimed.

This paper presents evidence to show that on one and the same day, two forecasters might make probability statements, each statement as valid as the other, and both forecasters equally skillful or performing with the same degree of accuracy or both. Yet one forecaster's probability statements might serve the operations better.

2. The general case

The *operational requirements* might be expressed by a table of scores (Gringorten, 1951). In table 1, it is assumed that there are n mutually exclusive categories of weather. For example n can equal two and the categories (X_1, X_2) can be *adverse* and *favorable* weather, respectively. There are (n^2) scores in the table. The score α_{ji} is the score awarded to the forecaster if his prediction is the event X_j but the subsequent is X_i. The score should be directly proportional to the profit resulting from the forecast, positive if there is a profit and negative if the operation incurs a net loss. For example, Thompson and Brier (1955) discuss the case of an operator who must decide whether or not to take protective measures against adverse weather. The cost of the protective measures is C, the loss suffered if no protective measures were taken in adverse weather is L and the income exclusive of loss or the cost of protection is T. In this case, there are four α's:

$$\left.\begin{aligned} \alpha_{11} &= T - C \\ \alpha_{12} &= T - C \\ \alpha_{21} &= T - L \\ \alpha_{22} &= T \end{aligned}\right\}. \tag{1}$$

The *climatology* of events might be summarized to show the frequency of each mutually exclusive event (X_i) following each initial condition (I_k) (table 2).

TABLE 1. Showing the score for each forecast or working assumption followed by each subsequent event.

Forecast or working assumption	Observed event X_1	X_i	X_n
X_1	α_{11}		α_{1n}
X_j		α_{ji}	
		α_{ii}	
X_n	α_{n1}		α_{nn}

TABLE 2. Illustrating a climatological summary of events. Initial conditions are symbolized by $I_1, \ldots, I_n$; subsequents by $X_1, \ldots, X_n$, and frequency of each combination by $P(I_kX_i)$.

Initial condition	Category of event X_1	X_i	X_n	Probability of initial condition
I_1	$P(I_1X_1)$		$P(I_1X_n)$	$P(I_1)$
I_k		$P(I_kX_i)$		$P(I_k)$
I_n	$P(I_nX_1)$		$P(I_nX_n)$	$P(I_n)$
Probability of event	$P(X_1)$	$P(X_i)$	$P(X_n)$	1.0

The symbol $P(I_kX_i)$ indicates the probability that the initial condition is I_k and the observed event is X_i. The symbol $P(I_k)$ denotes the climatic frequency (or probability) of I_k, $P(X_i)$ denotes the climatic frequency of X_i.

Two forecasters A and B might differ in their classifications of the weather. Let us suppose that A recognizes r categories of predictors $U_1, U_2, \cdots, U_r$ and B recognizes s categories of predictors $W_1, W_2, \cdots, W_s$. Let $P(U_lX_j)$ be the probability that the event X_j will occur among days classified as U_l, $P(W_LX_j)$ the probability that X_j will occur among days classified as W_L. If forecaster A forecasts perfectly then $P(U_lX_j)$ will be equal to $P(U_l)$ for one category and zero for all other categories of X.

But there is no perfect forecasting. Concerned as we are with the imperfect product of the meteorological art, we discuss in the following paragraphs three measures by which to evaluate the forecaster's efforts.

Measure of degree of accuracy.—Let us examine the problem of evaluating the accuracy of forecasters A and B. On the fraction of days $P(U_l)$, when forecaster A classifies his predictors as U_l there will be a fraction $P(U_lX_1)$ on which the subsequent weather is expected to be X_1. For a fraction $P(U_lX_2)$ the subsequent weather is expected to be X_2, and so on. If A chooses X_{jl} as his forecast on those days when the predictor classification is U_l, then the relative frequency with which he is expected to be correct is $P(U_lX_{jl})$. If $E(a_A)$ denotes the expected degree of accuracy for all of A's forecasts, then

$$E(a_A) = P(U_1X_{j1}) + \cdots + P(U_rX_{jr}) = \sum_{l=1}^{r} P(U_lX_{jl}). \quad (2)$$

Likewise, for forecaster B

$$E(a_B) = P(W_1X_{j1}) + \cdots + P(W_sX_{js}) = \sum_{L=1}^{s} P(W_LX_{jL}). \quad (3)$$

If forecaster A could forecast perfectly, each of his terms $P(U_lX_{jl})$ would be equal to $P(U_l)$ and $E(a_A)$ would become

$$\sum_{l=1}^{r} P(U_l) = 1.0.$$

Or, the perfect forecaster would obtain an average score of unity. The unskilled forecaster will obtain his maximum expected average score by forecasting the most frequent event, his average score being equal to the probability of that event (less than unity).

Measure of skill.—Thompson (1957) has summarized a viewpoint (Bross, 1953) on the process of improving the probability estimates by the addition of relevant information as that of increasing their "sharpness." Sharing this viewpoint we have used as our definition of *skill* (Gringorten, Lund, Miller, 1955): the forecaster's ability to analyze and classify the weather so that the probability of one subsequent event is estimated to increase more, in proportion to its climatic frequency, than the probability of any other subsequent event. This concept of skill has led us to a set of scores in which the score for a correctly predicted event is inversely proportional to the frequency of that event; the score for an incorrect forecast is nil. For a large number of unskilled forecasts the expected average score is a low figure (1.0), whether the unskilled system is a perpetual forecast of one event (*e.g.*, no rain) or whether it is always a forecast of persisting weather or whether it is a forecast by a system of chance, such as throwing darts at a target that has labelled areas "rain," "cloudy," and "clear" (see below).

To realize this goal of the scoring system the climatic frequencies of events have been determined subject to each initial condition. If I_k is the initial condition, X_j the future event, the *conditional climatic frequency* of X_j is

$$P(X_j|I_k)$$

and the score for correctly forecasting X_j is

$$1 \div P(X_j|I_k).$$

On the fraction of days $P(U_l)$ when forecaster A classifies his predictors as U_l there will be a smaller fraction $P(U_lI_kX_j)$ of days on which the initial condition is I_k and the subsequent weather is expected to be X_j. If, then, A's forecast on such days is X_j, the expected average score for the fraction $P(U_lI_k)$ of all days is

$$P(U_lI_kX_j)/P(X_j|I_k).$$

If $E(S_A)$ denotes A's expected average skill score for all of A's forecasts, then

$$E(S_A) = \sum_{l=1}^{r}\sum_{k=1}^{n} P(U_lI_kX_j)/P(X_j|I_k) = \sum_{l=1}^{r}\sum_{k=1}^{n} P(U_lI_kX_j)P(I_k)/P(I_kX_j). \quad (4)$$

Likewise, for B's forecasts,

$$E(S_B) = \sum_{L=1}^{s}\sum_{k=1}^{n} P(W_LI_kX_j)/P(X_j|I_k) = \sum_{L=1}^{s}\sum_{k=1}^{n} P(W_LI_kX_j)P(I_k)/P(I_kX_j). \quad (5)$$

For perfect forecasting, one or more U_l's would become identified with each predictand class X_j. There

would be, therefore, n groups of U_l's corresponding to the n classes of X, such that for each class, X_j

$$\sum_{l'} P(U_{l'}X_j) = P(X_j).$$

Also

$$\sum_{l'} P(U_{l'}X_j|I_k) = P(X_j|I_k).$$

Hence, it follows from equation (4) that, for perfect forecasting, $E(S_A)$ would become equal to the number of predictand classes, n.

The so-called "climatological" forecasts of a single event X_j reduce the probability of a correct forecast to the probability of X_j, causing the formula for $E(S_A)$ to reduce to

$$\sum_{k=1}^{n} P(I_kX_j)/P(X_j|I_k) = \sum_{k} P(I_k) = 1.0.$$

In "persistence" forecasting the only predictor is the initial class I_k. Therefore,

$$\sum_{l}\sum_{k} \rightarrow \sum_{k}$$

and $E(S_A)$ again reduces to unity (1.0).

For a system of chance or guesswork there is no relation between U_l and I_kX_j. Hence $E(S_A)$ again reduces to

$$\sum_{l}\sum_{k} P(U_l)P(I_k) = 1.0.$$

Therefore, the skill score defined by equation (4) above varies from a top expected value of n (the number of predictand classes) to an expected no-skill value of unity (1.0).

Measure of operational value.—The operational value of a forecast rests in its use as a working assumption. On each day that event out of the predictand set $X_1, \cdots, X_n$ should be chosen as the working assumption which will maximize the expected average profit.

In table 1 the score for each combination of forecast and outcome has been made proportional to the profit (or loss) resulting from the forecast or working assumption. If forecaster A recognizes the predictor condition as U_l, the frequency of which is $P(U_l)$, then the subsequent events will occur with frequencies,

$$P(U_lX_1), P(U_lX_2), \cdots, P(U_lX_n).$$

Let us suppose that X_j is chosen as a working assumption. Then the intended scores would be $\alpha_{j1}, \cdots, \alpha_{jn}$ for outcomes $X_1, \cdots, X_n$. The expected average score for the chosen forecast X_j is

$$\sum_{i=1}^{n} P(U_lX_i)\alpha_{ji}.$$

The working assumption or forecast X_j should be so chosen as to maximize this expression.

More generally, for each class U_l of the predictors, that X_j should be chosen which yields the maximum expected gain. Hence, where $E(G_A)$ is the expected gain from all of A's forecasts

$$E(G_A) = \sum_{l=1}^{r} \{\sum_{i=1}^{n} P(U_lX_i)\alpha_{ji}\}_{\max}, \qquad (6)$$

where $\{\ \}_{\max}$ denotes the X_j is chosen as the working assumption to yield the maximum value of the bracketed quantity.

Likewise

$$E(G_B) = \sum_{L=1}^{s} \{\sum_{i=1}^{n} P(W_LX_i)\alpha_{ji}\}_{\max}. \qquad (7)$$

If A could forecast perfectly, then, for each U_l he would forecast X_i with outcome inevitably X_i. The operational score $E(G_A)$ would become

$$\sum_{l=1}^{r} P(U_lX_i)\alpha_{ii} = \sum_{l=1}^{r} P(U_l)\alpha_{ii}.$$

That is, the expected perfect score would be the weighted average of the scores along the diagonal of table 1. The weights $P(U_l)$, of course, are dependent upon the climatology of the station.

For non-skilled forecasting $E(G_A)$ becomes

$$\sum_{l=1}^{r}\sum_{i=1}^{n} P(U_l)P(X_i)\alpha_{ji} = \sum_{l=1}^{r} P(U_l)\sum_{i=1}^{n} P(X_i)\alpha_{ji}$$

which expression differs from the previous expression for perfect forecasting by replacement of each α_{ii} with a weighted average of lesser values α_{ji}. If forecasts are unskilled the net profit can be maximized by perpetually forecasting that X_j for which the weighted average of the α_{ji}'s is largest.

Is it conceivable that a skillful predictor classification, U, will not serve a useful purpose to the operation? Yes, it is. (See chapter 4 below.)

The comparison of scores.—If we examine the three kinds of expected scores, $E(a)$, $E(S)$, and $E(G)$, we can find no reason to believe that the forecasts of A and B need to have equal operational value whenever they earn equal scores for accuracy or for skill or both. One forecaster may do better operationally simply by increasing the number of forecasts among those that pay the greatest profits or cause the least loss. The operationally better forecaster need not even display the greater degree of accuracy among the better paying events. But there has to be a greater number of profitable forecasts among which one or more α_{ji}'s are large. The following chapter is devoted to a special case to illustrate the argument of this paragraph.

3. Special case—dichotomous classifications

The above notation is for the generalized case and is admittedly cumbersome due largely to the summation signs. For dichotomous classifications of predictors, initial condition and predictand, the summation signs can be eliminated. Table 1 becomes table 3, table 2 becomes table 4 and the performances of forecasters A and B can be tabulated as in table 5.

TABLE 3. The scores for each combination of working assumption and subsequent when there are only two events to forecast.

Forecast	Score for each observation X_1	X_2
X_1	α_{11}	α_{12}
X_2	α_{21}	α_{22}

TABLE 4. The probabilities of each combination of initial condition and subsequent with marginal frequencies.

Initial condition	Observe X_1	X_2	Probability of initial condn.
I_1	$P(I_1X_1)$	$P(I_1X_2)$	$P(I_1)$
I_2	$P(I_2X_1)$	$P(I_2X_2)$	$P(I_2)$
Probability of subsequent	$P(X_1)$	$P(X_2)$	

TABLE 5. The probability of each predictor and subsequent.

A's predictor	Subsequent X_1	X_2	B's predictor	Subsequent X_1	X_2
U_1	$P(U_1X_1)$	$P(U_1X_2)$	W_1	$P(W_1X_1)$	$P(W_1X_2)$
U_2	$P(U_2X_1)$	$P(U_2X_2)$	W_2	$P(W_2X_1)$	$P(W_2X_2)$

In the dichotomous cases, the expressions (2) to (7) inclusive become

$$E(a_A) = P(U_1X_1) + P(U_2X_2) \tag{8}$$

$$E(a_B) = P(W_1X_1) + P(W_2X_2) \tag{9}$$

$$\begin{aligned} E(S_A) = {} & P(U_1I_1X_1)/P(X_1|I_1) \\ & + P(U_1I_2X_1)/P(X_1|I_2) \\ & + P(U_2I_1X_2)/P(X_2|I_1) \\ & + P(U_2I_2X_2)/P(X_2|I_2) \end{aligned} \tag{10}$$

$$\begin{aligned} E(S_B) = {} & P(W_1I_1X_1)/P(X_1|I_1) \\ & + P(W_1I_2X_1)/P(X_1|I_2) \\ & + P(W_2I_1X_2)/P(X_2|I_1) \\ & + P(W_2I_2X_2)/P(X_2|I_2) \end{aligned} \tag{11}$$

$$\begin{aligned} E(G_A) = {} & P(U_1X_1)\alpha_{11} + P(U_1X_2)\alpha_{12} \\ & + P(U_2X_1)\alpha_{21} + P(U_2X_2)\alpha_{22} \end{aligned} \tag{12}$$

$$\begin{aligned} E(G_B) = {} & P(W_1X_1)\alpha_{11} + P(W_1X_2)\alpha_{12} \\ & + P(W_2X_1)\alpha_{21} + P(W_2X_2)\alpha_{22}. \end{aligned} \tag{13}$$

From the above relations:

$$\begin{aligned} E(a_A) - E(a_B) = {} & \{P(U_1X_1) - P(W_1X_1)\} \\ & + \{P(U_2X_2) - P(W_2X_2)\} \end{aligned} \tag{14}$$

$$\begin{aligned} E(G_A) - E(G_B) = {} & (\alpha_{11} - \alpha_{21})\{P(U_1X_1) - P(W_1X_1)\} \\ & - (\alpha_{12} - \alpha_{22})\{P(U_2X_2) - P(W_2X_2)\}. \end{aligned} \tag{15}$$

If both systems of forecasting are *equally accurate*, then

$$\begin{aligned} E(G_A) - E(G_B) = {} & \{P(U_1X_1) - P(W_1X_1)\} \\ & \times \{(\alpha_{11} - \alpha_{21}) - (\alpha_{22} - \alpha_{12})\}. \end{aligned}$$

Supposing that $P(U_1X_1) > P(W_1X_1)$, then system A is more profitable than system B if $(\alpha_{11} - \alpha_{21}) > (\alpha_{22} - \alpha_{12})$ and system B is the more profitable if $(\alpha_{11} - \alpha_{21}) < (\alpha_{22} - \alpha_{12})$.

Example.—Suppose X_1, X_2 denote adverse and favorable weather, respectively, I_1, I_2 denote the initial conditions of adverse and favorable weather, U_1, W_1 denote the forecasts of adverse weather by A and B, respectively, and U_2, W_2 denote the forecasts of favorable weather. Let the terms in tables 4 and 5 have the values in table 6, and let the terms $P(U_lI_kX_j)$ have the values in table 7. Then

$$E(a_A) = E(a_B) = 0.8.$$

Also,

$$E(S_A) = E(S_B) = 19/12.$$

Therefore, both sets of forecasts are *equally skillful and equally accurate.* Yet

$$E(G_A) - E(G_B) = 0.1\{(\alpha_{11} - \alpha_{21}) - (\alpha_{22} - \alpha_{12})\}.$$

Or, at any time when the difference between scores for the successfully and unsuccessfully predicted event X_1 is greater than the difference between scores for the successfully and unsuccessfully predicted event X_2, then forecaster A is preferred to forecaster B. Forecaster A recognizes more cases as favoring X_1 although he is not as accurate in such forecasts.

TABLE 6. Example of probability figures to illustrate tables 4 and 5.

Initial condition	Relative frequency X_1	X_2	A's predictor class	Relative frequency X_1	X_2	B's predictor class	Relative frequency X_1	X_2
I_1	0.3	0.2	U_1	0.4	0.1	W_1	0.3	0
I_2	0.2	0.3	U_2	0.1	0.4	W_2	0.2	0.5

TABLE 7. Expansion of the example of table 6 to give values for the terms $P(U_lI_kX_j)$ and $P(W_iI_kX_j)$.

A's predictor class and initial condition	Relative frequency of observation X_1	X_2	B's predictor class and initial condition	Relative frequency of observation X_1	X_2
U_1I_1	0.2	0.1	W_1I_1	0.2	0
U_1I_2	0.2	0	W_1I_2	0.1	0
U_2I_1	0.1	0.1	W_2I_1	0.1	0.2
U_2I_2	0	0.3	W_2I_2	0.1	0.3

4. The critical ratio—dichotomy

Forecaster A might have a critical classification of the weather, U_c, such that the forecast of X_1 would net the same gain (or loss) to the operations as the forecast of X_2. That is:

$$P(U_cX_1)\alpha_{11} + P(U_cX_2)\alpha_{12} = P(U_cX_1)\alpha_{21} + P(U_cX_2)\alpha_{22}$$

from which

$$P(X_1|U_c) = P(U_cX_1)/P(U_c) = (\alpha_{22} - \alpha_{12})/\{(\alpha_{11} - \alpha_{21}) + (\alpha_{22} - \alpha_{12})\}.$$

As an example, let us use the α's of Thompson and Brier's (1955) paper (equations (1) above). Then

$$P(X_1|U_c) = C/L.$$

Thompson and Brier have referred to the conditional probability $P(X_1|U_c)$ as a *critical ratio*, equal in this instance to the "cost-to-loss" ratio. If, as in an earlier article by Thompson (1950), $C = 400$, $L = 5000$ then the critical ratio is 8 per cent. If, in table 6 above (not table 7), X_1 stands for rain, X_2 for no rain, then neither forecaster A nor B can do the operations any good, and precautionary measures against rain should always be taken. But if X_2 stands for rain and X_1 for no rain, then forecaster B can save the operations money on 30 per cent of the days.

5. Conclusions

Equally skillful and equally accurate sets of forecasts can serve a single operation with different success. In theory, at least, the issuing of one probability statement on a single day is not the most useful method to meet every operational requirement. If there is no method by which to combine the forecasts of A and B to obtain an improved combination of predictors then the operation should be a deciding factor in a choice between the forecasters or between the forecast systems or between two sets of probability statements.

REFERENCES

Bross, I. D. J., 1953: *Design for decision*, New York City, Macmillan Co., 272 pp.

Gringorten, I. I., 1951: The verification and scoring of weather forecasts. *J. Amer. Statistical Assn.*, **46**, 279–296.

——, I. A. Lund, and M. A. Miller, 1955: A program to test skill in terminal forecasting. *AF Survey in Geophys.*, No. 80, AFCRC, Bedford, Mass.

Thompson, J. C., 1950: A numerical method for forecasting rainfall in the Los Angeles area. *Mon. Wea. Rev.* **78**, 113–124.

——, and G. W. Brier, 1955: The economic utility of weather forecasts. *Mon. Wea. Rev.*, **83**, 249–254.

Thompson, J. C., 1957: Operations research looks at the weather forecast. *Proc. First Conf. Applied Meteor.*, Amer. Meteor. Soc. (Unpublished Manuscript) (9 pp.).