The Evaluation of Yes/No Forecasts for Scientific and Administrative Purposes

FRANK WOODCOCK

Bureau of Meteorology, Melbourne, Victoria, Australia

(Manuscript received 22 March 1976, in revised form 27 July 1976)

ABSTRACT

The basis upon which skill scores for evaluating yes/no categorical forecasts for scientific and administrative purposes depends is discussed and many of the common discriminants (formulas from which skill scores are derived) are reviewed and compared. The common process of subjecting forecasts to a trial consisting of a mixture of event and non-event occasions is outlined.

Those discriminants which prove to be measures of a forecasting technique's skill are shown, with the exception of Hanssen and Kuipers' (1965) discriminant, to give skill scores which depend upon the mixture of events and non-events in the trial. All these discriminants give incompatible rankings of forecasts because they are based on different standards of skill. It is shown that this discrepancy is resolved by ensuring that the trials under which forecasts are compared have equal numbers of event and non-event occasions; under these conditions, rankings become compatible. Hanssen and Kuipers' discriminant is shown to give the best estimate on an "unequal" trial to that expected if equalization were to be enforced. Hence, it is argued that Hanssen and Kuipers' discriminant is universally acceptable for evaluating yes/no forecasts for scientific and administrative purposes. Finally, the variance of Hanssen and Kuipers' discriminant is given to enable the statistical significance of the difference between two scores to be assessed and thereby make comparisons between techniques more meaningful.

1. Introduction

The importance of forecast evaluation for scientific, administrative and economic purposes (elaborated by Panofsky and Brier, 1958) has been widely accepted (Brier and Allen, 1952; Dobryshman, 1972). Controversy, however, surrounds the evaluation procedure and, until it is resolved, the usefulness of forecast evaluation is thereby diminished.

This paper is concerned with the evaluation of simple yes/no categorical forecasts for scientific and administrative purposes with a view to establishing a universally acceptable procedure. It is not intended to include economic (i.e., operational) considerations in the evaluation process because they require a knowledge of consumer strategy which is a consumer-dependent variable. Essentially this requires that successful and unsuccessful forecasts have different economic impacts on each consumer and therefore the evaluation procedure should weight the forecasts differently for each consumer. For scientific and administrative purposes it is accepted in the literature that an equal weighting of successes and failures is appropriate (Brier and Allen, 1952; Dobryshman, 1972).

The usual evaluation procedure is to subject the predictors to a series of trials consisting of a mixture of event and non-event days and to array the successes and failures in a 2×2 contingency table such as Table 1. The elements (A,B,C,D) are then manipulated according to a specified function, referred to here as discriminant, to yield skill scores which indicate relative accuracies of the predictors. By choosing the highest skill score the most accurate predictor relative to some standard predictor can then be obtained. There are four points worth emphasizing in this procedure. First, it is common practice to compare predictors under different trial conditions (i.e., different mixtures of event and non-event days). Schrank (1961), for example, compares two predictors, the first in its trial having a ratio of event days to non-event days of 1.23 and the second a ratio of 0.36. Second, a large part of the controversy in the literature surrounds acceptable definitions of accuracy and skill. Different discriminants often yield differing rankings of predictors because they are based on different definitions and the derived skill scores are measures of different attributes of the forecasts. Third, by compressing the elements (A,B,C,D) into a single number, there is inevitably a loss of information. Finally, since the selection of the best predictor involves comparing skill scores which are derived from samples of larger populations, it is necessary to compute their standard deviations in order to assess whether or not the differences between them are statistically significant.

2. Accuracy and the standard predictors

Supposing a standard predictor successfully partitions S_1 occurrences in a trial and another standard predictor successfully partitions S_2 occurrences in the

TABLE 1. Definition of the elements A, B, C, D of the 2×2 contingency table.

	Forecast	
Observed	Yes	No
Yes	A	B
No	C	D

same trial, then S_1-S_2 can be taken as the standard interval of accuracy for that trial. Any other predictor can be assessed for its accuracy in the same trial by the measured interval $S-S_2$, where S is the number of successful partitions gained by the predictor of interest. The skill of the predictor can be expressed as

$$\text{Skill}=\frac{S-S_2}{S_1-S_2}. \tag{1}$$

The skill of a predictor is therefore a function of (i) the standard predictors' successful partitions, (ii) the trial and (iii) the successful partitions gained by the predictor being evaluated. Discriminants, if they are to measure skill, should conform with the general formula (1). If the skill is a function of the trial conditions, it will be referred to as trial dependent.

The *perfect predictor* (P_p) is often used as one standard. This predictor is defined as successfully partitioning all occurrences into events and non-events in all trials.

If a trial consists of N occurrences, then

$$\left.\begin{aligned} S_p&=A_p+D_p=N\\ B_p&=C_p=0 \end{aligned}\right\}, \tag{2}$$

where the suffix p indicates a perfect predictor process.

The *false predictor* (P_f) is that predictor which fails to successfully partition any of the occurrences in all trials. Thus

$$\left.\begin{aligned} S_f&=A_f+D_f=0\\ B_f+C_f&=N \end{aligned}\right\}, \tag{3}$$

where the suffix f indicates a false predictor process.

The *random predictor* (P_r) is that predictor which partitions occurrences as if forecasts and occurrences were stochastically independent in all trials. Under this condition

$$\left.\begin{aligned} A_r&=(A+B)(A+C)/N\\ B_r&=(B+A)(B+D)/N\\ C_r&=(C+D)(C+A)/N\\ D_r&=(D+C)(D+B)/N \end{aligned}\right\}, \tag{4}$$

where the suffix r indicates a random predictor process.

The *best unskilled predictor* (P_u) was introduced by Appleman (1960). It is defined as that predictor which correctly partitions all the events or all the non-events, whichever is the larger category in the trial, and fails to correctly partition any of the other category. Should both categories be equally represented, then one of them is successfully partitioned and the other completely incorrectly partitioned. Hence, if $(A+B)>(C+D)$,

$$\left.\begin{aligned} S_u&=A_u=A+B\\ B_u&=D_u=0\\ C_u&=C+D \end{aligned}\right\}, \tag{5}$$

or if $(C+D)>(A+B)$,

$$\left.\begin{aligned} S_u&=D_u=C+D\\ C_u&=A_u=0\\ B_u&=A+B \end{aligned}\right\}, \tag{6}$$

where the suffix u indicates a best unskilled predictor process. With equal categories either distribution (A_u,B_u,C_u,D_u) can be used.

3. Review of some common meteorological discriminants

a. Ratio test (R)

The Ratio test is the simplest of the common discriminants and is defined as the ratio of successful partitions to the total number of occurrences in the trial. Thus

$$R=\frac{A+D}{N}, \tag{7}$$

where $0\leqslant R\leqslant 1$. R can be rewritten in terms of two of the standard predictors as

$$R=\frac{S-S_f}{S_p-S_f}. \tag{8}$$

This formulation clearly shows that R uses a standard interval based on the perfect and false predictors; the measured interval expressed as a ratio of this standard interval indicates the skill according to (1). P_p scores 1 and P_f scores 0 in all trials, while P_r can score anything between 0 and 1 and P_u anything between 0.5 and 1.

b. Skill test (S_k)

The Skill test is defined as twice the ratio of the number of successful partitions in excess of those scored by the random predictor to the total number of occurrences in the trial. In terms of contingency table elements this reduces to

$$S_k=\frac{4(AD-BC)}{N^2}, \tag{9}$$

where $-1\leqslant S_k\leqslant 1$. The factor 2 in the definition is

merely to expand the scale to convenient limits of ± 1; its presence does not affect the ranking of predictors. The factor $(AD-BC)$ in the numerator appears frequently in discriminants; wherever it occurs the random predictor scores zero in all trials.

In terms of standard predictors

$$S_k=\frac{2(S-S_r)}{S_p-S_f}. \tag{10}$$

The Skill test can be seen to depend upon the perfect and false predictors for its standard interval and the random forecast for its measured interval and therefore does not measure skill in accordance with Eq. (1). P_p scores 1 in the trial where $A=D=N/2$ but generally scores between 0 and 1, while P_f scores -1 when $B=C=N/2$ but generally scores between 0 and -1. P_u scores 0 in all trials.

c. Heidke skill score (T) (Brier and Allen, 1952)

This discriminant has been widely used in the United States and is defined as the ratio of the excess of successful partitions over those gained by the random predictor, to the excess of successful partitions gained by the perfect predictor over those gained by the random predictor. Hence

$$T=\frac{S-S_r}{S_p-S_r}. \tag{11}$$

The definition clearly shows that skill is defined in terms of P_p and P_r. P_p scores 1 in all trials and P_r scores 0 as indeed does P_u. However, scores attained by P_f are trial dependent and can vary between -1 when $B=C=N/2$ and 0 when B or C are 0.

In term of contingency table elements

$$T=\frac{2(AD-BC)}{(A+B)(B+D)+(A+C)(C+D)}, \tag{12}$$

where $-1\leqslant T\leqslant 1$.

d. Appleman's (1960) discriminant (U)

This is analogous to Heidke's skill score except that the best unskilled predictor replaces the random predictor. Here

$$U=\frac{S-S_u}{S_p-S_u}. \tag{13}$$

P_p and P_u are used to define accuracy with P_u scoring 0 in all trials and P_p scoring 1; P_f and P_r both yield trial dependent scores. In the case where $(A+B)>(C+D)$,

$$U=\frac{D-B}{C+D}, \tag{14}$$

where $-B/C\leqslant U\leqslant 1$. When $(C+D)>(A+B)$, U is determined by interchanging B and C and replacing D with A. If $(A+B)=(C+D)$, either form can be used and $-1\leqslant U\leqslant 1$.

e. Hanssen and Kuipers' (1965) discriminant (V)[1]

This discriminant is essentially the sum of two ratio tests. The trial is considered in two parts, one relating to the event occasions and the other to the non-event occasions. V can be expressed as

$$V=\left(\frac{S-S_f}{S_p-S_f}\right)_{\text{events}}+\left(\frac{S-S_f}{S_p-S_f}\right)_{\text{non-events}}-1. \tag{15}$$

The minus one term is merely a scale shift and does not affect the ranking of predictors. V can be seen to define accuracy in terms of P_p and P_f which score 1 and -1, respectively, in all trials. Additionally, it can be seen that the P_u must score 0 in all trials since it must score 1 and 0 in the two ratio tests. In contingency table elements

$$V=\frac{AD-BC}{(A+B)(C+D)}, \tag{16}$$

where $-1\leqslant V\leqslant 1$. The numerator here ensures that P_r scores zero in all trials, so that the V discriminant yields scores for all the standard predictors which are independent of the trial conditions.

f. Schrank's (1961) discriminant (W)

This was devised to overcome differences in ranking of predictors between the Ratio test and Heidke's skill score. It has been shown here that these two discriminants differ in their definitions of accuracy and thus skill and hence it is not surprising that they sometimes produce differing results. Schrank, although he did not state it, was forced to redefine accuracy.

In its simplest form, Schrank's discriminant can be written

$$W=\frac{R+S-1}{2}, \tag{17}$$

where $-1\leqslant W\leqslant 0.5$. In terms of standard predictor successes this can be rewritten

$$W=\frac{S-0.5S_r}{S_p-S_f}-0.5. \tag{18}$$

Here S_p and S_f define the standard interval and the measured interval is relative to half the successful partitions gained by the random predictor. All the standard predictors yield trial-dependent scores, P_p scores between 0 and 0.5, P_f scores between -1 and

[1] Also independently derived by Dobryshman (1972).

-0.5, P_r between -0.5 and 0 and P_u between -0.25 and 0.

g. The correlation coefficient (r)

This is derived from the application of the usual statistical linear correlation coefficient between two variables. If (X,Y) are the variables representing the observed and the forecast occasions, respectively, where each observed or forecast occurrence is given a value of 1 and each observed or forecast non-occurrence is given a value 0, then the correlation coefficient

$$r=\frac{N\Sigma(XY)-\Sigma(X)\Sigma(Y)}{\{[N\Sigma(X^2)-(\Sigma X)^2][N\Sigma(Y^2)-(\Sigma Y)^2]\}^{\frac{1}{2}}} \quad (19)$$

reduces to

$$r=\frac{AD-BC}{[(A+B)(A+C)(C+D)(B+D)]^{\frac{1}{2}}}, \quad (20)$$

where $-1\leqslant r\leqslant 1$. P_r scores 0 in all trials, P_p scores 1 and P_f scores -1, while P_u scores are indeterminate.

It is interesting to note that the least-squares linear regression straight line between observed and forecast occurrences has a slope of $(AD-BC)/(A+B)(C+D)$ which is Hanssen and Kuipers' discriminant. Furthermore, if the outcome of the trial were such that D equalled A or C equalled B then r would equal V.

g. Other discriminants

Several other discriminants have been mentioned in the literature and will be briefly discussed here for the sake of completeness. None of them is suitable for forecast evaluation and they will not be considered beyond this section.

1) The chi-square test (χ^2), when applied to Table 1, reduces to

$$\chi^2=\frac{N(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)}, \quad (21)$$

where $\chi^2\geqslant 0$. P_r scores 0 in all trials and P_u scores are indeterminate, while P_p and P_f have trial-dependent scores. The normal procedure is to evaluate χ^2 and to ascertain its significance from standard tables with one degree of freedom. The difficulty with the use of χ^2 is that N occurs in the numerator and only a large enough sample needs to be chosen for almost any result to be highly significant. A fuller discussion of this aspect of the χ^2 test is given in many books on elementary statistics (e.g., Wonnacott and Wonnacott, 1972).

2) Two other discriminants, Yule's Q (Moroney, 1963), where

$$Q=\frac{AD-BC}{AD+BC}, \quad (22)$$

with $-1\leqslant Q\leqslant 1$, and Yule's Y (Moroney, 1963), where

$$Y=\frac{\sqrt{AD}-\sqrt{BC}}{\sqrt{AD}+\sqrt{BC}}, \quad (23)$$

with $-1\leqslant Y\leqslant 1$, may also be encountered in the literature. These discriminants prove unsuitable for forecast evaluation since they only require B or C to be zero in order to yield maximum scores, whereas it should be essential for B and C to be zero together for a maximum score to be obtained.

The main features of the discriminants deemed acceptable for forecast evaluation at this stage are shown in Table 2.

4. The effect of trial conditions

A further insight into the differences between the discriminants discussed can be gained by considering the scores obtained by a specified predictor under different trial conditions. Suppose, for example, a particular technique (P_1) is able to successfully predict the occurrence of thunderstorms, except whenever the 800–500 mb shear exceeds some critical value which, on the long-term average, occurs on 25% of thunderstorm days. Furthermore, suppose the method successfully predicts non-thunderstorm days except when some other criterion is exceeded which, on the long-term average, occurs on 50% of non-thunderstorm days. If 11 different representative samples, each of 200 days, are chosen from all days with the only difference between the samples being the ratio of thunderstorm to non-thunderstorm days, then a table such as Table 3 could result.

The most important conclusion to be drawn from Table 3 is that the skill scores generally are trial dependent. This implies that the common practice of subjecting different forecasting techniques to trials containing different trial conditions and selecting or ranking them in order of skill on the basis of the skill scores attained from a common discriminant can be grossly misleading, simply because the skill so measured is not solely a function of the predictor but is also a function of the trial conditions.

Furthermore, with the aid of Table 3, it can be demonstrated that another common practice, namely that of evaluating two forecast techniques over the same trial, can be misleading. By considering another specified predictor (P_2) which successfully partitions 50% of the events and 75% of the non-events and subjecting it to trial 4 (i.e., line 4 of Table 3), it would gain a partitioning of (70, 70, 15, 45) and the various skill scores it would attain correspond to those in line

TABLE 2. Main properties of common meteorological discriminants. T. D. indicates trial dependency, other abbreviations are explained in the text.

Discriminant		Range Min	Range Max	Standard predictor scores P_p	P_f	P_r	P_u	Interval Standard	Interval Measured
Ratio	(R)	0	1	1	0	T.D.	T.D.	S_p-S_f	$S-S_f$
Skill	(S_k)	−1	1	T.D.	T.D.	0	0	S_p-S_f	$S-S_r$
Heidke	(T)	−1	1	1	T.D.	0	0	S_p-S_r	$S-S_r$
Appleman	(U)	$-B/C$	1	1	T.D.	T.D.	0	S_p-S_u	$S-S_u$
Hanssen and Kuipers	(V)	−1	1	1	−1	0	0	S_p-S_f	$S-S_f$
Schrank	(W)	−1	0.5	T.D.	T.D.	T.D.	T.D.	S_p-S_f	$S-(S_r/2)$
Correlation coefficient	(r)	−1	1	1	−1	0	—	—	—

8 of Table 3. Here all the discriminants bar the Skill test and Hanssen and Kuipers' discriminant would rate P_1 more skillful than P_2. However, if the two predictors were subjected to trial 8, P_1 would score as line 8 and P_2 as line 4 and hence P_2 would be generally rated more skillful than P_1.

These considerations show that the skill scores of most of the common discriminants are trial dependent and that this dependency essentially means that in any trial wherein the events and non-events are not equally represented is a biased trial. Hence, to overcome the bias, it is necessary to either use Hanssen and Kuipers' discriminant, which is not trial dependent, or to use a standardized trial in which the events and non-events are equally represented. A further point in favor of having such a standardized trial is that it is precisely the condition required to overcome the conflict in the definitions of accuracy adopted by the different discriminants.

Under equalization of event and non-event occurrences, P_r and P_u score 0.5 in the ratio test; wherever P_f scores were trial dependent in Table 2 they now score -1; P_p scores $+1$ in all discriminants except Schrank's in which it scores 0.5. In Schrank's discriminant both P_u and P_r score -0.25 and the range of Appelman's discriminant is symmetrical about zero with $-1 \leqslant U \leqslant 1$. With all the standard predictors' scores being constants through the equalization of events and non-events in the trial, it can be shown that

$$S=T=U=V=2R-1 \tag{24}$$

and

$$W=\frac{3R}{2}-1. \tag{25}$$

Hence, all the discriminants (except r) become equivalent in their ranking of predictors. The relationships (24) and (25) can be seen in the distribution (75, 25, 50, 50) of Table 3. Under equality of events and non-events r reduces to

$$r=\frac{N(2R-1)}{[N^2-4(A-D)^2]^{\frac{1}{2}}}. \tag{26}$$

For any fixed value of R, the minimum value of r occurs whenever $A=D$. As there does not seem to be any justification for a predictor to be considered less accurate when this occurs, r is unacceptable as a discriminant.

TABLE 3. Variations of discriminant skill scores for a forecast which successfully partitions 75% of events and 50% of non-events with varying trial conditions.

Contingency table element A	B	C	D	Discriminant skill scores Ratio	Skill	Heidke	Appleman	Hanssen and Kuipers	Schrank	Correlation coefficient
150	50	0	0	0.750	0.000	0.000	—	—	−0.125	—
135	45	10	10	0.725	0.090	0.141	−1.750	0.250	−0.093	0.168
120	40	20	20	0.700	0.160	0.211	−0.500	0.250	−0.070	0.218
105	35	30	30	0.675	0.210	0.244	−0.087	0.250	−0.058	0.245
90	30	40	40	0.650	0.240	0.255	0.125	0.250	−0.055	0.257
75	25	50	50	0.625	0.250	0.250	0.250	0.250	−0.063	0.258
60	20	60	60	0.600	0.240	0.231	0.000	0.250	−0.080	0.250
45	15	70	70	0.575	0.210	0.198	−0.417	0.250	−0.108	0.232
30	10	80	80	0.550	0.160	0.151	−1.250	0.250	−0.145	0.201
15	5	90	90	0.525	0.090	0.087	−3.750	0.250	−0.193	0.150
0	0	100	100	0.500	0.000	0.000	—	—	−0.25	—

5. Theoretical and random equalization of event and non-event days, significance tests and Hanssen and Kuipers' discriminant

In practice, the recommended procedure would be to list separately the event and non-event days and then the elements of the contingency table (A,B,C,D) could be obtained by random selection from each of the lists ensuring that when the selection is completed $(A+B)$ equals $(C+D)$. From the distribution (A,B,C,D) thus obtained the discriminant $(2R-1)$ could be evaluated, thereby providing an unbiased and acceptable measure of the predictor's accuracy. However, this procedure is fairly cumbersome and it is worth investigating whether Hanssen and Kuipers' discriminant, when applied to the original trial conditions, gives a good estimate of $(2R-1)$ under equalization.

Suppose in a particular trial $(A+B)>(C+D)$; then, in order to equalize the event and non-event categories, a sample of size $(C+D)$ is required to be drawn from the population $(A+B)$. The number a of successful forecasts in the different possible samples of size $(C+D)$ drawn from $(A+B)$ is a hypergeometric variable and its probability distribution $P(a)$ is given by

$$P(a)=\frac{\binom{A}{a}\binom{B}{C+D-a}}{\binom{A+B}{C+D}}, \tag{27}$$

where $\binom{A}{a}$ indicates the number of different samples of size a that can be drawn from population A, etc. The mathematical expectation of a, $[E(a)]$, is

$$E(a)=\frac{A(C+D)}{A+B}. \tag{28}$$

Of the number b of failures, the expectation $E(b)$ is

$$E(b)=\frac{B(C+D)}{A+B}. \tag{29}$$

Hence the expected proportion of expected successes is

$$\frac{E(a)}{E(a)+E(b)} \quad \text{which is} \quad \frac{A}{A+B};$$

i.e., the same as in the original trial. It follows, therefore, that the value V from the original trial

$$V=\frac{A}{A+B}+\frac{D}{C+D}-1 \tag{30}$$

is the same as the expected value after random selection to enforce equalization.

Hanssen and Kuipers' discriminant, therefore, must be regarded as providing the best estimate of $(2R-1)$ under equalization. They derived the variance (σ_V^2) of V as

$$\sigma_V^2=\frac{N^2-4(A+B)(C+D)V^2}{4N(A+B)(C+D)}, \tag{31}$$

which enables skill scores to be compared more meaningfully. If V is close to 1 or, if sample sizes are small, the upper bound on the confidence limits may exceed 1, in which case 1 is taken instead. Such practice is not uncommon (e.g., Moroney, 1963).

6. Conclusion

Different discriminants used in the literature have been reviewed. Chi-square, the correlation coefficient, and Yule's Q and Y discriminants have been shown to be unacceptable for forecast evaluation purposes. The Ratio test, the Skill test, Heidke's skill score, Appleman's discriminant, and Hanssen and Kuipers' and Schrank's discriminants have been shown to produce incompatible rankings of forecasts because they are based on different standards of accuracy. Additionally, with the exception of Hanssen and Kuipers' discriminant, they yield, in general, trial-dependent skill scores.

These two problems can be resolved if the trial conditions are equalized, with respect to events and non-events. All the discriminants give consistent rankings and $(2R-1)$ is an acceptable discriminant under these conditions. Hanssen and Kuipers' discriminant, when applied to the trial, gives the best estimate of $(2R-1)$ that could be attained if the trial results were to be randomly equalized. Hence, their discriminant provides an unbiased and acceptable measure of forecast accuracy for scientific and administrative purposes.

Acknowledgment. This paper is published by permission of the Director of Meteorology, Australia.

REFERENCES

Appleman, H. S., 1960: A fallacy in the use of skill scores. *Bull. Amer. Meteor. Soc.*, **41**, 64–67.

Brier, G. W., and R. A. Allen, 1952: Verification of weather forecasts. *Compendium of Meteorology*, Boston, Amer. Meteor. Soc., 841–848.

Dobryshman, E. M., 1972: Review of forecast verification techniques. WMO Tech. Note, No. 120, 17–20.

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. Koninklijk Nederlands Meteorologisch Institut, *Meded. Verhand.*, **81**, 2–15.

Moroney, M. J., 1963: *Facts from Figures*. Penguin Books Ltd., 264–266.

Panofsky, H. A., and G. W. Brier, 1958: *Some Applications of Statistics to Meteorology*. Pennsylvania State University Press, 191–194.

Schrank, W. L., 1961: A solution to the problem of evaluating forecast techniques. *Bull. Amer. Meteor. Soc.*, **42**, 277–280.

Wonnacott, T. H., and R. J. Wonnacott, 1972: *Introductory Statistics*, 2nd ed. Wiley, 423–428.