

NOTES AND CORRESPONDENCE

On Reducing Probability Forecasts to Yes/No Forecasts

IAN MASON

A.C.T. Regional Office, Australian Bureau of Meteorology, Canberra City, Australia

17 March 1978 and 12 October 1978

ABSTRACT

Some scores for yes/no forecasts discussed in a paper by Woodcock (1976) are further considered. An expression for a probability p_0 is found for any score such that the expected value of the score is maximized if the event is forecast when its probability is greater than p_0 and not forecast if its probability is less than p_0 . Particular expressions for p_0 for the scores discussed by Woodcock are presented. Asymptotic values are mostly near either 0.5 or the sample relative frequency of the event. Comments are made on the dependence of p_0 on sample size and order of verification. The relation of this note to earlier work by Bryan and Enger and also Thompson and Brier is briefly discussed.

1. Introduction

In a recent article Woodcock (1976) has reviewed some scoring rules for the evaluation of yes/no forecasts and commented on the effect of varying trial conditions on ranking by values of scores. This note is essentially a footnote to Woodcock's, from the point of view of a forecaster who must issue yes/no forecasts and who wishes to make use of his knowledge of the uncertainty in the meteorological situation to maximize his score. The scores discussed below and the order of their presentation follow Woodcock's paper.

Weather forecasts are usually uncertain to some degree, and valid information about the amount of uncertainty in a forecast can have economic value if used rationally (see Thompson and Brier, 1955; Thompson, 1963; Mason 1975). There is now adequate evidence that experienced weather forecasters can make meaningful subjective assessments of the uncertainty in their forecasts (Sanders, 1963, 1973; Stael von Holstein, 1971; Murphy and Winkler, 1974). The loss of information, and hence of value, when probabilistic forecasts are converted to yes/no form has been discussed in the published literature by, for example, Thompson (1962, 1971), Mason (1976) and Murphy (1977). Efficient yes/no forecasts can be given if the forecaster has enough knowledge of the particular operation in which his forecasts are to be applied to make reasonably confident estimates of certain economic parameters (Thompson and Brier, 1955; Murphy, 1977). Without this knowledge the least misleading way of issuing fore-

casts is in general to give them explicitly as probabilities.

However, it appears that many forecasters and users of weather forecasts are at present either unable or unwilling to make use of probabilistic information and hence that predictions will for some time continue to be given as yes/no statements. If forecasts of this kind are to be verified using scoring rules, it will be interesting to know the value of the probability p_0 , say, that maximizes the expected value of the score if yes is forecast whenever the probability of the event is greater than p_0 , and no when it is less. This note derives a general expression for p_0 and gives particular expressions for the scores discussed by Woodcock.

Bryan and Enger (1967) presented strategies for converting a set of probability forecasts into categorical forecasts so as to maximize the "asymptotic" values of three scores; the Heidke, Vernon and Appleman skill scores. The problem they addressed was that of comparing the accuracy of yes/no forecasts with the accuracy of probability forecasts in situations involving any number of events. The main difference between their study and the present one is that this note finds the appropriate tactic, on a day-to-day basis, by which a forecaster can maximize the expected value of the next increment to his score. Bryan and Enger prescribe a rule which does not vary from day to day, and focus attention on the long-run consequences of adhering to this rule. Also, this note considers only the two-state situation. The value of p_0 for the Heidke score implied by their analysis is very close for large samples

TABLE 1. Possible values of the score S after the $(N + 1)$ th forecast has been verified. Note that positive orientation implies $S_{11} > S_{10}$ and $S_{00} > S_{01}$.

		Forecast	
		Yes	No
Observed	Yes	S_{11}	S_{01}
	No	S_{10}	S_{00}

to the exact value found using the expression derived below. The Vernon score is not considered in this note. In the case of the Appleman score the two methods give the same result.

2. A general expression for p_0

Consider a score for categorical forecasts whose value is represented by S . It can be assumed without loss of generality that the score has a positive orientation, i.e., that greater values indicate closer correspondence between forecasts and observations. Suppose that after N forecasts have been verified the value of the score is S_N . The forecaster must now choose a forecast (either yes or no) for the $(N + 1)$ th occasion. The possible values of S_{N+1} can be arranged in a 2×2 table (Table 1), where S_{ij} is the value of S_{N+1} if category i is forecast and category j occurs.

The expected value of S_{N+1} if yes is forecast, as a function of the probability p of the event is

$$E_1(S_{N+1}) = pS_{11} + (1 - p)S_{10}, \tag{1}$$

where the subscript 1 on E_1 denotes expected value when yes is forecast.

It can be shown that $E_1(S_{N+1})$ is a monotonic increasing function of p on the interval $[0,1]$ so long as S has positive orientation. This is plausible since for a forecast that the event will occur the expected value of the score should increase as the probability of the event increases.

For $p = 0, E_1(S_{N+1}) = S_{10}$ and for $p = 1, E_1(S_{N+1}) = S_{11}$, using (1).

Therefore, since $S_{11} > S_{10}$ (positive orientation), the value of $E_1(S_{N+1})$ at $p = 0$ is less than the value at $p = 1$.

Differentiating (1) with respect to p gives

$$\frac{\partial}{\partial p} [E_1(S_{N+1})] = S_{11} - S_{10} > 0 \tag{2}$$

since S has positive orientation. Thus, $E_1(S_{N+1})$ is monotonic increasing on the interval $[0,1]$.

Similarly, if no is forecast, then

$$E_0(S_{N+1}) = pS_{01} + (1 - p)S_{00} \tag{3}$$

and by the same reasoning as above $E_0(S_{N+1})$ is a monotonic decreasing function of p on $[0,1]$.

For a reasonable scoring rule it is also necessary that both $S_{00} > S_{10}$ and $S_{11} > S_{01}$. If $S_{11} > S_{01}$ and $S_{10} > S_{00}$, then $E_1(S_{N+1}) > E_0(S_{N+1})$ for all p —and one may always expect a higher score for a forecast of yes regardless of its probability. Similarly, if $S_{01} > S_{11}$ and $S_{00} > S_{10}$, then $E_0(S_{N+1}) > E_1(S_{N+1})$ for all p and one should never forecast the event.

It follows that there is a point p_0 in $[0,1]$ for which $E_1(S_{N+1}) = E_0(S_{N+1})$.

Also, if $p < p_0$ then $E_0(S_{N+1}) > E_1(S_{N+1})$ and if $p > p_0$, then $E_1(S_{N+1}) > E_0(S_{N+1})$.

Thus, if $p < p_0$ the expected value of the score is greater if no is forecast, and if $p > p_0$ a forecast of yes has the larger expected value.

A general expression for p_0 can be found from (1) and (3) above as follows:

$$\text{At } p = p_0, E_1(S_{N+1}) = E_0(S_{N+1}). \tag{4}$$

Substituting from (1) and (3) gives

$$p_0S_{11} + (1 - p_0)S_{10} = p_0S_{01} + (1 - p_0)S_{00}, \tag{5}$$

so that

$$p_0 = \frac{S_{00} - S_{10}}{S_{11} - S_{01} + S_{00} - S_{10}}. \tag{6}$$

This expression yields the familiar decision rule involving the cost-loss ratio (Thompson and Brier, 1955) when the scores are replaced by appropriate costs and losses, i.e., $S_{10} = S_{11} = C, S_{00} = 0, S_{01} = L$.

3. p_0 for some scores

Following Woodcock (1976) the results of a series of forecasts to be scored are presented in Table 2.

a. Ratio score (Woodcock, 1976; also, Brier and Allen, 1952, where it is called "percent correct")

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

where A, D and N are verification table elements (Table 2), $F = A + D$, the number of correct forecasts and R stands for the value of the ratio score.

TABLE 2. Result of verification of a series of forecasts. A, B, C and D are numbers of forecasts in each category.

		Forecast		
		Yes	No	Total
Observed	Yes	A	B	$A + B$
	No	C	D	$C + D$
	Total	$A + C$	$B + D$	$N = A + B + C + D$

Using the notation of Section 2, we have

$$F_{00} = F_{11} = \frac{F_N + 1}{N + 1}, \tag{8}$$

$$F_{01} = F_{10} = \frac{F_N}{N + 1}. \tag{9}$$

Substituting in (6) gives $p_0 = 0.5$. Thus a forecaster who knows that his predictions are to be evaluated using the ratio score should forecast the event whenever its probability is greater than 0.5, and forecast nonoccurrence if it is less than 0.5.

The value of p_0 for this score does not depend either on the previous value of the score or on the order in which the forecasts are verified.

b. Skill test (Woodcock, 1976)

$$S = \frac{4(AD - BC)}{N^2}, \tag{10}$$

where S stands for the value of the score and A, B, C, D and N are defined by Table 2.

For this score,

$$p_0 = \frac{A + B}{N}. \tag{11}$$

This is just the sample relative frequency of the event in the first N forecasts. A forecaster being assessed with this score should give yes as his forecast if the probability of the event is greater than its relative frequency in the set of previously verified forecasts and no if the probability is less. If he does not know the current value of the sample relative frequency, then his best tactic would be to use for p_0 the climatological probability of the event.

c. Heidke score (Brier and Allen, 1952)

$$H = \frac{F - E}{N - E}, \tag{12}$$

where H stands for the value of the score, F is the number of correct forecasts in the sample of N , and E is the number expected correct based on some standard such as chance, persistence or climatology. It has been common practice to follow Brier and Allen (1952, p. 846) and take for E the expected value of the number of correct forecasts using the sample relative frequency as the probability of the event with the sample frequency of forecasts, i.e., in terms of the elements of Table 2,

$$E = \frac{(A + B)}{N}(A + C) + \frac{(C + D)}{N}(B + D). \tag{13}$$

Appleman (1960) pointed out that when E is calculated in this way the Heidke score does not reliably indicate which of two competing techniques is more

accurate, and a positive value does not necessarily imply that a skilled forecast procedure is superior to an unskilled procedure. Schrank (1961) also criticized this score. However, it is apparently still used.

The dependence of E as defined above on sample values makes the expression for p_0 very difficult to simplify. Eq. (6) above could be used as it stands, i.e.,

$$p_0 = \frac{H_{00} - H_{10}}{H_{11} - H_{01} + H_{00} - H_{10}}, \tag{14}$$

where H_{ij} is the value of the Heidke score after the next, $(N + 1)$ th, forecast has been verified if i is forecast and j occurs. Bryan and Enger (1967) find a decision rule for this score [Eq. (2.15) in their paper] by using climatological probabilities in the expression for E and neglecting certain terms which are relatively small for large N . In the case of two categories only their criterion can be written.

$$p_0 = p_c(1 - H) + H/2, \tag{15}$$

where p_c is the climatological probability of the event and H is an "optimum" value for the Heidke score found by iteration. In practice if (15) is used with sample relative frequency substituted for p_c and the current value of the score for H , then values p_0 are found which are very close to those given by the exact expression (14) above.

If the forecaster knows beforehand whether the comparison method forecasts an occurrence of the event or not, for example, when persistence is used to give the value of E , then the following expressions can be found for p_0 for this score:

1) When the comparison method forecasts occurrence,

$$p_0 = \frac{N - E}{2(N - E) + 1}, \tag{16}$$

which tends to 0.5 from below as $N - E$ tends to infinity.

2) When the comparison method forecasts non-occurrence,

$$p_0 = \frac{N - E + 1}{2(N - E) + 1}, \tag{17}$$

which tends to 0.5 from above as $N - E$ tends to infinity. In both (16) and (17) the difference $p_0 - 0.5$ is less than 0.05 for $N - E > 5$.

Thus for practical purposes, so long as the comparison method has had more than five failures in the set of forecasts so far verified and E is found by a method which gives a definite forecasts of yes or no on each occasion, forecasters being verified with the Heidke score should use $p_0 = 0.5$.

It should be noted, however, that if E is found using the sample marginal frequencies [Eq. (13)], then 0.5 will not necessarily be the best value for

p_0 , which can be calculated using either Eq. (14) (exact) or (15) (approximate).

d. *Appleman's score* (Appleman, 1960)

$$U = \frac{F - X}{N - X}, \quad (18)$$

where U stands for the value of the score, F the number of correct forecasts and X the number of observations in the more frequently observed category, i.e.,

$$X = \begin{cases} A + B, & \text{if } (A + B) > (C + D) \\ C + D, & \text{if } (C + D) > (A + B) \end{cases} \quad (19)$$

$X = A + B$ leads to

$$p_0 = \frac{C + D}{2(C + D) + 1}, \quad (20)$$

which tends to 0.5 from below as $C + D$ tends to infinity. The difference $p_0 - 0.5$ is less than 0.05 for $(C + D) > 5$. $X = C + D$ leads to

$$p_0 = \frac{A + B + 1}{2(A + B) + 1}, \quad (21)$$

which tends to 0.5 from above as $A + B$ tends to infinity. Again, the difference $p_0 - 0.5$ is less than 0.05 for $(A + B) > 5$.

Thus for practical purposes forecasters being evaluated with the Appleman score could use $p_0 = 0.5$ if the number of observations in the less frequently occurring class is greater than 5.

Bryan and Enger (1967) found that to maximize the asymptotic value of the Appleman score for a forecast in any number of categories one should forecast the category with the highest "true" probability. For two categories this is equivalent to forecasting the event if its probability is greater than 0.5, in accordance with the result obtained above.

e. *Hanssen and Kuipers' score* (Hanssen and Kuipers, 1965)

$$V = \frac{P_{00}}{P_0} + \frac{P_{11}}{P_1} - 1. \quad (22)$$

where $P_{00} = D/N$, $P_{11} = A/N$, $P_0 = (C + D)/N$ and $P_1 = (A + B)/N$. With these expressions substituted, Eq. (22) can be written

$$V = \frac{AD - BC}{(A + B)(C + D)}. \quad (23)$$

For this score

$$p_0 = \frac{A + B + 1}{N + 2}. \quad (24)$$

This tends to the sample relative frequency of

the event as N increases. For $N > 20$ the difference between p_0 and the sample relative frequency is less than 0.05.

f. *Schrank's score* (Schrank, 1961)

$$S = \frac{F - (R + E)}{N}, \quad (25)$$

where S is the value of the score, F the number of correct forecasts, R one-half of the number of wrong forecasts and E is as defined by (13).

Schrank's score may be expressed in terms of p_0 by

$$p_0 = \frac{N(1 + 4f) + 1}{2(3N + 1)}, \quad (26)$$

where $f = (A + C)/N$, the sample relative frequency of forecasts of the event.

Note that p_0 for this score cannot be less than 0.17 ($f = 0$, $N \rightarrow \infty$) or greater than 0.83 ($f = 1$, $N \rightarrow \infty$).

4. Comments

Except for the ratio score, all the scores considered in Section 3 give expressions for p_0 which depend on sample frequencies in the set of forecasts already verified. Thus, p_0 varies from one occasion to the next, and also depends on the order in which the forecasts are verified. In practice day-to-day fluctuations in p_0 are negligible for sample sizes ≥ 20 . For smaller samples exact values for p_0 can easily be calculated.

The dependence on order of verification is more of a problem. Unless the forecaster knows the order in which his forecasts will be verified his only rational tactic is to use the "asymptotic" values for p_0 given above. This could be quite seriously in error. If, for example, Woodcock's (1976, p. 1213) suggestion is adopted for a standardized trial in which events and non-events are equally represented, then the sample relative frequency of the event is clearly 0.5. If a score is used for which the asymptotic value of p_0 is the climatological relative frequency then categorical forecasts issued on this basis will not maximize the expected value of the score unless, fortuitously, the climatological relative frequency is 0.5. Ideally, of course, the forecaster is fully informed of the scoring method.

Thompson and Brier (1955) speculated on the best value for p_0 for public weather forecasts where a wide range of operations is involved. They discussed, from the point of view of maximizing the economic value of the skill over climatology demonstrated by the forecasts, two possible values for p_0 , namely, 0.5 and the climatological probability of the event. There are reasonable arguments for both values which are briefly as follows. Setting

$p_0 = 0.5$ results in prediction of the event with about the same frequency as it occurs; this is generally regarded as desirable. However, this alone does not ensure that the resulting decisions are economically sound. Thompson and Brier point out that where the event of interest is infrequent but of considerable consequence it may be preferable to use p_0 near the climatological probability. The number of false alarms will be high but may still be acceptable if losses when the event occurs unforecast are large relative to the cost of protection.

It is interesting that the forecaster who is trying to maximize his score on one of the rules discussed in this note (except for the Heidke score when E is calculated using marginal frequencies and Schrank's score) will be obliged to select for p_0 one of the two values discussed by Thompson and Brier. The administrator who wishes to use one of these scores will therefore need to decide which tactic he wants to encourage in his forecasters and can choose a scoring rule accordingly.

Acknowledgment. This note is published by permission of the Director of Meteorology.

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*, Amer. Meteor. Soc., 841–848.
- Bryan, J. G., and I. Enger, 1967: Use of probability forecasts to maximize various skill scores. *J. Appl. Meteor.*, **6**, 762–769.
- Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Korink. Neder. Meteor. Inst., Meded. Verhand.*, **81**, 3–15.
- Mason, I. B., 1975: The economic value of probability weather forecasts. Tech. Rep. No. 16, Australian Bureau of Meteorology, 10 pp.*
- Mason, I. B., 1976: An example of the difference between the values of categorical and probability forecasts. *Meteor. Note*, No. 86, Australian Bureau of Meteorology, 8 pp.*
- Murphy, A. H., 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816.
- , and R. L. Winkler, 1974: Credible interval temperature forecasting: Some experimental results. *Mon. Wea. Rev.*, **102**, 784–794.
- Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.
- , 1973: Skill in forecasting daily temperatures and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.*, **54**, 1171–1179.
- Schrank, W. R., 1961: A solution to the problem of evaluating forecast techniques. *Bull. Amer. Meteor. Soc.*, **42**, 277–280.
- Staël von Holstein, C. A. S., 1971: An experiment in probabilistic weather forecasting. *J. Appl. Meteor.*, **10**, 635–645.
- Thompson, J. C., 1962: Economic gains from scientific advances and operational improvements in meteorological predictions. *J. Appl. Meteor.*, **1**, 13–17.
- , 1963: Weather decision making—The pay-off. *Bull. Amer. Meteor. Soc.*, **44**, 75–78.
- , 1971: Probability, decision models and the value of improved weather forecasts. *Preprints 3rd Conf. Int. Symp. Probability and Statistics in Applied Sciences*, Honolulu, Amer. Meteor. Soc., 90–93.
- , and G. W. Brier, 1955: The economic utility of weather forecasts. *Mon. Wea. Rev.*, **83**, 249–253.
- Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.*, **104**, 1209–1214.

* Available from Australian Bureau of Meteorology, P.O. Box 1289K, Melbourne.