Hedging and Skill Scores for Probability Forecasts

ALLAN H. MURPHY

National Center for Atmospheric Research, Boulder, Colo. 80302

(Manuscript received 20 June 1972, in revised form 20 October 1972)

ABSTRACT

An individual skill score (SS) and a collective skill score (CSS) are examined to determine whether these scoring rules are strictly proper or improper. The SS and the CSS are both standardized versions of the Brier, or probability, score (PS) and have been used to measure the “skill” of probability forecasts. The SS is defined in terms of individual forecasts, while the CSS is defined in terms of collections of forecasts. The SS and the CSS are shown to be improper scoring rules, and, as a result, both the SS and the CSS encourage hedging on the part of forecasters.

The results of a preliminary investigation of the nature of the hedging produced by the SS and the CSS indicate that, while the SS may encourage a considerable amount of hedging, the CSS, in general, encourages only a modest amount of hedging, and even this hedging decreases as the sample size \( K \) of the collection forecasts increases. In fact, the CSS is approximately strictly proper for large collections of forecasts \((K \geq 100)\).

Finally, we briefly consider two questions related to the standardization of scoring rules: 1) the use of different scoring rules in the assessment and evaluation tasks, and 2) the transformation of strictly proper scoring rules. With regard to the latter, we identify standardized versions of the PS which are strictly proper scoring rules and which, as a result, appear to be appropriate scoring rules to use to measure the “skill” of probability forecasts.

1. Introduction

Meteorologists have often used the Brier, or probability, score \((PS)\) (Brier, 1950), a measure of the “accuracy” of individual forecasts (see Murphy and Winkler, 1970, 1971), to evaluate probability forecasts. The PS is also a strictly proper scoring rule (e.g., Murphy and Epstein, 1967; Winkler and Murphy, 1968), and, as a result, the PS discourages hedging on the part of forecasters.\(^2\)

The results of empirical studies indicate that the “accuracy” of precipitation probability forecasts, as measured by the PS, is related to the climatological probability of precipitation (see, for example, Sanders, 1963; Hughes, 1965; Glahn and Jorgenson, 1970). Specifically, for climatological probabilities less than one-half, the “accuracy” of the forecasts increases, in general, as the climatological probability increases, while for climatological probabilities greater than one-half, their “accuracy” decreases, in general, as the climatological probability increases. In order to measure the “accuracy” of these forecasts relative to climatology, “skill scores” have been formulated in terms of the PS for the forecasts and the PS for the climatological probabilities (see Sanders, 1963; Hughes, 1965; Glahn and Jorgenson, 1970). Specifically, an “individual” skill score (SS) has been defined in terms of individual forecasts and a “collective” skill score (CSS) has been defined in terms of collections of forecasts. The CSS is presently used by the National Weather Service (NWS) to measure the “skill” of their precipitation probability forecasts (Hughes, 1965; Glahn and Jorgenson, 1970).\(^3\)

The purposes of this paper are to show that the SS and the CSS are not strictly proper scoring rules, to present the results of a preliminary investigation of the nature of the hedging produced by the SS and the CSS, and to briefly consider two questions related to the standardization of scoring rules.

In Section 2 we define the PS, the SS, and the CSS. The concepts of hedging and strictly proper and improper scoring rules are briefly discussed in Section 3. In Sections 4 and 5 we show that the SS and the CSS, respectively, are not, in general, strictly proper scoring rules. We present the results of a preliminary investigation of the nature of the hedging produced by the SS and the CSS in Section 6. The results indicate that while the SS may encourage a considerable amount of hedging, the CSS, in general, encourages only a modest amount of hedging, and even this hedging decreases as the sample size of the collection of forecasts increases. Two questions related to the standardization of scoring rules are briefly considered in Section 7: 1) the use of

\(^1\) The National Center for Atmospheric Research is sponsored by the National Science Foundation.

\(^2\) For a recent detailed description of the nature and properties of the PS, refer to Murphy (1970).

\(^3\) The need for such standardized scoring rules, in order to be able to compare forecasters, forecast offices, etc., has recently been emphasized by Hughes (1971).
different scoring rules in the assessment and evaluation tasks, and 2) the transformation of strictly proper scoring rules. With regard to the latter, we identify standardized versions of the PS which are strictly proper scoring rules. Section 8 consists of a brief summary and conclusion.

2. The individual and collective skill scores

a. The PS

Consider a meteorological variable whose range has been divided into $N$ mutually exclusive and collectively exhaustive states $\{s_1, \cdots, s_N\}$. Let the vector

$$r = (r_1, \cdots, r_N) \quad (r_n \geq 0, \sum_{n=1}^{N} r_n = 1 \; n = 1, \cdots, N)$$

denote the forecaster’s forecast, where $r_n$ is the forecast probability of state $s_n$, and let the vector $d = (d_1, \cdots, d_N)$ denote the relevant observation, where $d_n = 1$ if state $s_n$ obtains and $d_n = 0$ otherwise.

Then, the PS for the forecast $r$ is defined as follows:

$$PS(r, d) = \sum_{n=1}^{N} (r_n - d_n)^2. \quad (1)$$

The range of the PS is the closed interval $[0, 2]$ and the PS has a negative orientation (i.e., the smaller the score the better the score). When state $s_j$ obtains $PS(r, d)$ becomes $PS_j(r)$, where

$$PS_j(r) = 1 - 2r_j + \sum_{n=1}^{N} r_n^2. \quad (2)$$

Now, consider a collection of $K$ forecasts $r_k = (r_{1k}, \cdots, r_{Nk})$ and the $K$ relevant observations $d_k = (d_{1k}, \cdots, d_{Nk})$ $(k = 1, \cdots, K)$. Then, the total PS for the collection of $K$ forecasts is $TPS(r, d)$, where, from (1),

$$TPS(r, d) = \sum_{k=1}^{K} \sum_{n=1}^{N} (r_{nk} - d_{nk})^2. \quad (3)$$

b. The SS

Let the vector

$$\pi = (\pi_1, \cdots, \pi_N) \quad (\pi_n \geq 0, \sum_{n=1}^{N} \pi_n = 1 \; n = 1, \cdots, N)$$

denote the set of relevant climatological probabilities, where $\pi_n$ is the climatological probability of state $s_n$.

Then, the individual skill score $SS$ for the forecast $r$ is defined as

$$SS(r, \pi, d) = \left[ PS(\pi, d) - PS(r, d) \right] / PS(\pi, d),$$

or

$$SS(r, \pi, d) = 1 - \left[ PS(r, d) / PS(\pi, d) \right]. \quad (4)$$

The range of the SS is the half-open interval $(-\infty, 1]$ and the SS has a positive orientation (i.e., the larger the score the better the score).\footnote{Note that the SS represents the fractional amount by which the PS for the forecast $(r_1, \cdots, r_N)$ improves upon the PS for the climatological probabilities $(\pi_1, \cdots, \pi_N)$. When state $s_j$ obtains, $SS(r, \pi, d)$ becomes $SS_j(r, \pi)$, where $SS_j(r, \pi) = 1 - \left[ PS_j(r) / PS_j(\pi) \right]$.}

The total SS for the collection of $K$ forecasts is $TSS(r, \pi, d)$, where, from (4),

$$TSS(r, \pi, d) = K - \sum_{k=1}^{K} \left[ PS(r_k, d_k) / PS(\pi_k, d_k) \right]. \quad (5)$$

The range of the SS is also the half-open interval $(-\infty, 1]$ and the SS has a positive orientation.\footnote{Note that the CSS represents the fractional amount by which the PS for the K forecasts $r_k$ improves upon the PS for the K climatological probabilities $\pi_k$ $(k = 1, \cdots, K)$. In addition, note that $CSS(r, \pi, d)$ is not, in general, equivalent (i.e., linearly related) to $TSS(r, \pi, d)$.\footnote{In fact, $CSS(r, \pi, d)$ is equivalent to $TSS(r, \pi, d)$ only if $PS(\pi, d) = PS(\pi, d_k)$ for all $j$ and $k(j, k = 1, \cdots, K; j \neq k)$, in which case $CSS(r, \pi, d) = (1/K) TSS(r, \pi, d)$.}}

The collective skill score CSS for the collection of $K$ forecasts is defined as

$$CSS(r, \pi, d) = \left[ \sum_{k=1}^{K} PS(\pi_k, d_k) - \sum_{k=1}^{K} PS(r_k, d_k) \right] / \sum_{k=1}^{K} PS(\pi_k, d_k),$$

or

$$CSS(r, \pi, d) = 1 - \left[ \sum_{k=1}^{K} PS(r_k, d_k) / \sum_{k=1}^{K} PS(\pi_k, d_k) \right]. \quad (7)$$

The range of the CSS is also the half-open interval $(-\infty, 1]$ and the CSS has a positive orientation.\footnote{Note that the CSS represents the fractional amount by which the PS for the $K$ forecasts $r_k$ improves upon the PS for the $K$ climatological probabilities $\pi_k$ $(k = 1, \cdots, K)$. In addition, note that $CSS(r, \pi, d)$ is not, in general, equivalent (i.e., linearly related) to $TSS(r, \pi, d)$.\footnote{In fact, $CSS(r, \pi, d)$ is equivalent to $TSS(r, \pi, d)$ only if $PS(\pi, d) = PS(\pi, d_k)$ for all $j$ and $k(j, k = 1, \cdots, K; j \neq k)$, in which case $CSS(r, \pi, d) = (1/K) TSS(r, \pi, d)$.}}

3. Hedging and strictly proper and improper scoring rules

Let the vector

$$p = (p_1, \cdots, p_N) \quad (p_n \geq 0, \sum_{n=1}^{N} p_n = 1 \; n = 1, \cdots, N)$$

denote the forecaster’s “degree of belief,” or judgment, where $p_n$ is the judgmental probability that state $s_n$ will obtain. Hedging is said to occur whenever a forecaster’s forecast $r$ does not correspond to his judgment $p$, i.e., whenever $r \neq p$.

A scoring rule, $S(r, d)$, with a positive orientation is said to be strictly proper if and only if the forecaster
maximizes his (subjective) expected score \( ESS(r,p) \), where

\[
ESS(r,p) = \sum_{j=1}^{N} p_j S_j(r),
\]

by making his forecast \( r \) correspond to his judgment \( p \) (e.g., Murphy and Epstein, 1967; Winkler and Murphy, 1968). A strictly proper scoring rule, then, discourages hedging on the part of forecasters. On the other hand, the scoring rule \( S(r,d) \) is said to be improper if the forecaster’s expected score \( S(r,p) \) attains a unique maximum for some forecast \( r \) which does not correspond to his judgment \( p \). An improper scoring rule, then, encourages hedging on the part of forecasters.

4. Hedging and the individual skill score

The SS, when state \( s_j \) obtains, becomes, from (2) and (5),

\[
SS_j(r,\pi) = 1 - \left[ (1 - 2r_j + \sum_{n=1}^{N} r_n)^{2}/PS_j(\pi) \right].
\]

Then, the expected score is \( ESS(r,\pi,p) \), where

\[
ESS(r,\pi,p) = \sum_{j=1}^{N} p_j SS_j(r,\pi),
\]
or, from (9),

\[
ESS(r,\pi,p) = 1 - \sum_{j=1}^{N} p_j \left[ (1 - 2r_j + \sum_{n=1}^{N} r_n)^{2}/PS_j(\pi) \right].
\]

Now, taking the derivative of \( ESS(r,\pi,p) \) with respect to \( r_m \), we obtain

\[
\frac{\partial ESS(r,\pi,p)}{\partial r_m} = \left[ 2p_m/PS_m(\pi) \right]
\]

\[
- (2r_m) \sum_{n=1}^{N} \left[ p_n/PS_n(\pi) \right],
\]

and setting this derivative equal to zero yields

\[
p_m - r_m^* PS_m(\pi) \sum_{n=1}^{N} \left[ p_n/PS_n(\pi) \right] = 0,
\]
or

\[
r_m^* = \left[ p_m/PS_m(\pi) \right] / \sum_{n=1}^{N} \left[ p_n/PS_n(\pi) \right],
\]

\[(m = 1, \ldots, N), (11)\]

where \( r_m^* \) is the “optimal” value of \( r_m \), i.e., the value of \( r_m \) which maximizes\(^6\) the forecaster’s expected SS for particular values of the climatological probabilities \( (\pi_1, \ldots, \pi_N) \) and the forecaster’s judgments \( (p_1, \ldots, p_N) \). Note that \( r_m^* \neq p_m \) unless \( \pi_m = 1/N \) for all \( n (n = 1, \ldots, N) \).

\(^6\) We assume that the forecaster’s utility function is linearly related to the scoring rule \( S(r,d) \). That is, we assume that the forecaster attaches a “value” to the score he receives for his forecast which is linearly related to the score itself.

\(^*\) Note that \( \partial ESS(r,\pi,p)/\partial r_m^2 < 0 \).

5. Hedging and the collective skill score

The concepts of hedging and strictly proper (and improper) scoring rules relate to individual probability forecasts and to a forecaster’s behavior when making such forecasts. Specifically, in order to determine whether a scoring rule is strictly proper or improper, we must assume that the forecaster has a definite strategy in mind when he makes a forecast. In this paper we shall assume that the forecaster has already made \( K \) forecasts \( r_k (k = 1, \ldots, K) \), that he wants to maximize his expected score on the \((K+1)\)st forecast \( r_{K+1} \), and that the \((K+1)\)st forecast is independent of the previous \( K \) forecasts.\(^9\)

The CSS, when state \( s_j \) obtains on the \((K+1)\)st occasion, becomes, from (3) and (7),

\[
CSS_j(r_{K+1},\pi_{K+1}) = \frac{TPS(r,d) + PS_j(r_{K+1})}{TPS(\pi,d) + PS_j(\pi_{K+1})}.
\]

Then, the expected CSS is \( ECSS(r,\pi,p) \), where

\[
ECSS(r,\pi,p) = \sum_{j=1}^{N} p_j CSS_j(r_{K+1},\pi_{K+1}),
\]
or, from (2) and (12),

\[
ECSS(r,\pi,p) = 1 - \sum_{j=1}^{N} p_j \left[ TPS(\pi,d) + PS_j(\pi_{K+1}) \right] \times \sum_{j=1}^{K+1} \left[ TPS_j(\pi,d) + PS_j(\pi_{K+1}) \right].
\]

Now, taking the derivative of \( ECSS(r,\pi,p) \) with respect to \( r_{m,K+1} \), we obtain

\[
\frac{\partial ECSS(r,\pi,p)}{\partial r_{m,K+1}} = \left[ 2p_m, K+1 / [TPS(\pi,d) + PS_m(\pi_{K+1})] \right]
\]

\[
- (2r_{m,K+1}) \sum_{n=1}^{N} \left[ p_n, K+1 / [TPS(\pi,d) + PS_n(\pi_{K+1})] \right].
\]

Setting this derivative equal to zero yields

\[
p_m, K+1 - r_{m,K+1} \sum_{n=1}^{N} \left[ p_n, K+1 / [TPS(\pi,d) + PS_n(\pi_{K+1})] \right] = 0,
\]

\[^9\] Other strategies are, of course, available to the forecaster. For example, we could have assumed that the forecaster, prior to making any forecasts, wants to maximize his expected score on each of \( K \) occasions and that the forecasts on the \( K \) occasions are dependent. Dynamic programming could be used to solve this problem for the set of \( K \) optimal forecasts \( r^*_k (k = 1, \ldots, K) \).
Table 1. The forecast probability $r^*_{1}$ which maximizes the forecaster's expected $SS$ in the two-state situation for selected values of the climatological probability $\pi$ and the forecaster's judgment $p_1$.

<table>
<thead>
<tr>
<th>Climatological probability $\pi$</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecaster's judgment $p_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.6</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.7</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.8</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1.0</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

† The value of $r^*_{1}$ is not defined for this combination of values of $\pi$ and $p_1$.

or

$$r^*_{m,K+1} = \{p_{m,K+1}/[TPS(\pi,d) + PS_n(\pi_{K+1})]\}$$

$$\sum_{n=1}^{N} \{p_{n,K+1}/[TPS(\pi,d) + PS_n(\pi_{K+1})]\}$$

(m = 1, ··· , N), (14)

where $r^*_{m,K+1}$ is the optimal value of $r_{m,K+1}$, note that $r^*_{m,K+1} \neq p_{m,K+1}$ unless $\pi_{n,K+1} = 1/N$ for all $n (n = 1, ··· , N)$ or unless $p_{m,K+1} = 0$ or 1. Thus, the CSS is not, in general, a strictly proper scoring rule.

6. The nature of the hedging produced by the individual and collective skill scores

a. Hedging and the $SS$

With regard to the nature of the hedging produced by the $SS$, we can identify several specific questions of interest:

1) How much hedging is produced by the $SS$, i.e., how large are the differences in probability between the optimal forecast $r^*$ and the judgment $p_1$?

2) How "flat" is the expected $SS$ function, $ESS(r, \pi, p_1)$, e.g., how large are the differences in expected score between the optimal forecast $r^*$ and the judgment $p_1$?

3) How does the flatness of $ESS(r, \pi, p_1)$ compare with that of the expected $PS$ function, $EPS(r, p_1)$, e.g., is the $SS$ more or less likely to encourage hedging than the $PS$ is to discourage hedging?

The amount of hedging produced by the $SS$ can be measured in terms of the difference in probability between the optimal forecast $r^*$ and the judgment $p_1$. In the two-state ($N = 2$), i.e., "precipitation-no precipitation," situation the optimal forecast is $r^* = (r^*_1, r^*_2)$, where, from (11),

$$r^*_1 = p_1/[(\pi_1 + (\pi_2/\pi_1)\pi_2)]$$

(15)

($r^*_2 = 1 - r^*_1$, $\pi_2 = 1 - \pi_1$, and $p_2 = 1 - p_1$). We present the

Note that $\partial ESS(r, \pi, p_1)/\partial r^*_{m,K+1} < 0$. 

![Fig. 1. The expected SS, ESS(r, \pi, p_1), in the two-state situation for selected values of the forecaster's judgment p_1 when the climatological probability \pi equals 0.1, (a), and 0.4, (b).](image-url)
values of $r_1^*$ in Table 1 for selected values of $\pi_1$ and $p_1$. Note that $r_1^* \neq \pi_1$ unless $\pi_1 = 0.5$ or unless $p_1 = 0.0$ or 1.0. In addition, note that the amount of hedging is quite large when the climatological probability $\pi_1$ is either greater than 0.6 or less than 0.4. For example, $|r_1^* - \pi_1| > 0.1$ for $0.2 \leq \pi_1 \leq 0.8$ when $\pi_1 \geq 0.6$ or $\pi_1 \leq 0.4$, while $r_1^* \leq (\leq) 0.9(0.1)$ for all values of $p_1$ when $\pi_1 \geq (\leq) 0.9(0.1)$. Specifically, if $\pi_1 = 0.2$ and $p_1 = 0.4$, then $r_1^* = 0.4$, while if $\pi_1 = 0.6$ and $p_1 = 0.8$, then $r_1^* = 0.9$. Thus, if a forecaster is aware that his forecasts are being evaluated by the SS and if his utility function is linearly related to the SS (see Footnote 7), then the SS could produce a considerable amount of hedging.

The amount of hedging which is produced by the SS also depends, in part, upon the flatness of the expected SS function, $ESS(r, \pi, p)$. For example, if this function is quite flat over a range of values which includes the optimal forecast $r^*$ as well as the judgment $p$, then the forecaster has little incentive to hedge and he may well make his forecast $r$ correspond closely to his judgment $p$. The function $ESS(r, \pi, p)$ in the two-state (i.e., precipitation—no precipitation) situation is depicted in Figs. 1a and 1b for selected values of $p_1$ when $\pi_1$ equals 0.1 and 0.4, respectively. Note that in a "dry" climate, i.e., when $\pi_1 = 0.1$, the function $ESS(r, \pi, p)$ is certainly not flat. Moreover, while the range of this function decreases as $p_1$ increases, the loss in expected score if the forecaster makes his forecast $r_1$ correspond to his judgment $p_1$ instead of maximizing the function by setting $r_1$ equal to the optimal forecast $r_1^*$, i.e., $ESS(r^*, \pi, p) - ESS(p, \pi, p)$, increases from 0.877 when $p_1 = 0.1$ to 11.899 when $p_1 = 0.8$. On the other hand, in a relatively "wet" climate, i.e., when $\pi_1 = 0.4$, the function $ESS(r, \pi, p)$ is quite flat (note the change in the scale on the ordinate). In addition, the range of $ESS(r, \pi, p)$ decreases as $p_1$ increases and the loss in expected score, $ESS(r^*, \pi, p) - ESS(p, \pi, p)$, increases from 0.016 when $p_1 = 0.1$ to only 0.089 when $p_1 = 0.8$. Thus, the shape of the function $ESS(r, \pi, p)$ is such that the SS encourages a considerable amount of hedging on the part of forecasters, particularly when the climatological probability $\pi_1$ is small.  

The loss in expected score, $ESS(r^*, \pi, p) - ESS(p, \pi, p)$, in the two-state situation is presented in Table 2 for selected values of $\pi_1$ and $p_1$.

In order to compare the SS and the PS with regard to the degree to which the former encourages and the latter discourages hedging, the expected scores must be transformed in such a way that the range and orientation of $ESS(r, \pi, p)$ and $EPS(r, p)$ are identical. We present these transformed expected scores, $ESS^*(r, \pi, p)$ and $EPS^*(r, p)$, in the two-state situation in Figs. 2a and 2b, respectively, for selected values of $p_1$ when $\pi_1$ equals 0.1 and 0.4. A comparison of $ESS^*(r, \pi, p)$ and $EPS^*(r, p)$ is difficult because these functions attain their maxima, and occasionally even their minima, for different values of $r_1$. However, with reference to their respective maxima, the function $ESS^*(r, \pi, p)$ appears to be slightly flatter, i.e., less "sensitive," than the function $EPS^*(r, p)$. Thus, in terms of relative expected scores, the SS is slightly less likely to encourage hedging than the PS is to discourage hedging. The reader should be aware that the discussion in this paragraph has been concerned with relative expected scores, while the discussion in the previous paragraph was concerned with absolute expected scores.

In this regard, the absolute sensitivity of a scoring rule can be changed by performing a linear transformation on the scoring rule; on the other hand, such a transformation will not change its relative sensitivity. However, the use of such a transformation may produce a scoring rule whose range is such that the forecaster's

\[ ESS(r^*, \pi, p) - ESS(p, \pi, p), \]

\[ 0.0 \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4 \quad 0.5 \quad 0.6 \quad 0.7 \quad 0.8 \quad 0.9 \quad 1.0 \]

\[ \begin{array}{cccccccccccc}
  0.0 & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\
  0.1 & 0.000 & 0.877 & 3.112 & 6.113 & 9.288 & 12.044 & 13.792 & 13.938 & 11.899 & 7.111 & 0.000 \\
  0.2 & 0.000 & 0.197 & 0.692 & 1.348 & 2.025 & 2.585 & 2.893 & 2.818 & 2.250 & 1.139 & 0.000 \\
  0.3 & 0.000 & 0.065 & 0.227 & 0.432 & 0.634 & 0.782 & 0.836 & 0.762 & 0.547 & 0.226 & 0.000 \\
  0.4 & 0.000 & 0.016 & 0.056 & 0.102 & 0.143 & 0.166 & 0.167 & 0.139 & 0.089 & 0.031 & 0.000 \\
  0.5 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
  0.6 & 0.000 & 0.031 & 0.089 & 0.139 & 0.167 & 0.166 & 0.143 & 0.102 & 0.056 & 0.016 & 0.000 \\
  0.7 & 0.000 & 0.226 & 0.547 & 0.762 & 0.836 & 0.782 & 0.634 & 0.432 & 0.227 & 0.065 & 0.000 \\
  0.8 & 0.000 & 1.139 & 2.250 & 2.818 & 2.893 & 2.585 & 2.025 & 1.348 & 0.692 & 0.197 & 0.000 \\
  0.9 & 0.000 & 7.111 & 11.899 & 13.938 & 13.792 & 12.044 & 9.288 & 6.113 & 3.112 & 0.877 & 0.000 \\
  1.0 & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\
\end{array} \]

\[ \uparrow \text{The loss in expected score, } ESS(r^*, \pi, p) - ESS(p, \pi, p), \text{ is not defined for } \pi_1 = 0.0 \text{ or } 1.0. \]

\[ \text{This statement is equally valid when the climatological probability } \pi_1 \text{ is large. For example, } ESS(r, \pi, p) \text{ when } \pi_1 = 0.9 \text{ is a mirror image of } ESS(r, \pi, p) \text{ when } \pi_1 = 0.1, \text{ and, in addition, the loss in expected score, } ESS(r^*, \pi, p) - ESS(p, \pi, p), \text{ when } \pi_1 = 0.9 \text{ and } p_1 = 0.2 \text{ is the same as that when } \pi_1 = 0.1 \text{ and } p_1 = 0.8. \]

\[ \text{A different linear transformation is required for each combination of values of } \pi_1 \text{ and } p_1 \text{ for the SS. For example, when } \pi_1 = 0.1 \text{ and } p_1 = 0.8, \text{ the function } ESS(r, \pi, p) \text{ is } 0.952 \text{ when } p_1 = 0.8 \text{ and } 0.997 \text{ when } p_1 = 0.1. \]

\[ \text{The function } ESS^*(r, \pi, p) \text{ attains its maximum at } r_1 = r_1^* \text{ and its minimum at } r_1 = 0 \text{ if } p_1 \geq 0.5 \text{ and at } r_1 = 1 \text{ if } p_1 < 0.5, \text{ while the function } EPS^*(r, p) \text{ attains its maximum at } r_1 = p_1, \text{ its minimum at } r_1 = 0 \text{ if } p_1 \geq 0.5 \text{ and at } r_1 = 1 \text{ if } p_1 < 0.5. \]
utility function is no longer linear in the score (see Footnote 7). For example, a forecaster’s utility function may be linear in SS(ψ,d) but not in SSS∗(ψ,d) (or vice versa). Thus, while linear transformations of scoring rules provide a means of changing the absolute flatness of expected scoring rule functions, the use of such transformations may introduce nonlinear utility considerations.

b. Hedging and the CSS

Since the CSS is defined in terms of a collection of forecasts, in order to investigate the nature of the hedging produced by the CSS we must consider collections of forecasts rather than individual forecasts. In this paper we shall be primarily concerned with the amount of hedging produced by the CSS for hypothetical collections of forecasts in the two-state situation.

The optimal forecast in the two-state situation is

\[ r_{K+1}^∗ = (r_{1,K+1}^∗, r_{2,K+1}^∗) \]

where, from (14),

\[ r_{1,K+1}^∗ = \frac{\{p_{1,K+1}/[TPS(ψ,d) + π_{2,K+1}]\} \{p_{1,K+1}/[TPS(ψ,d) + π_{2,K+1}]\}}{\{p_{1,K+1}/[TPS(ψ,d) + π_{2,K+1}]\} + \{p_{2,K+1}/[TPS(ψ,d) + π_{2,K+1}]\} + \{p_{2,K+1}/[TPS(ψ,d) + π_{2,K+1}]\}} \]

Note that \( r_{1,K+1}^∗ \) depends upon the PS for the K climatological probabilities \( π_k \) \( (k=1, \ldots, K) \), \( TPS(ψ,d) \), but not upon the PS for the K forecasts \( r_k \) \( (k=1, \ldots, K) \), \( TPS(r,d) \). Thus, we are actually concerned not with the forecasts in the hypothetical collections but with the climatological probabilities which correspond to the forecasts.

Consider a collection of ten forecasts \((K=10)\) in the two-state situation and suppose that the climatological probabilities of states \( s_1 \) and \( s_2 \) are equal to 0.2 and 0.8, respectively, on all of the ten occasions \([i.e., π_{1k} = 0.2 \text{ and } π_{2k} = 0.8 \text{ (} k=1, \ldots, K) \] Further, suppose that states \( s_1 \) and \( s_2 \) obtain on two and eight of the ten occasions, respectively \([i.e., d_1 = 0.2 \text{ and } d_2 = 0.8 \text{; } d_{1k} = (1/K) \sum d_{nk} (n=1,2; k=1, \ldots, K) \] Then, \( TPS(ψ,d) \) for this collection of forecasts is, from (3), 0.26. Note that the amount of hedging produced by the CSS, 0.04, is considerably smaller than the amount of hedging that would have been produced by the SS, 0.36 \((r_{1}^∗ = 0.4 \text{ when } π_{1}^∗ = 0.2 \text{ and } π_{2}^∗ = 0.4 \); see Table 1). Moreover, even this modest amount of hedging decreases as the sample size \( K \) of the collection of forecasts increases. For example, if the forecaster had made forecasts on twenty occasions \((K=20)\) with the same climatological probabilities, i.e., \( π_{1k} = 0.2 \text{ and } π_{2k} = 0.8 \text{ (} k=1, \ldots, K) \), and if states \( s_1 \) and \( s_2 \) had obtained on four and sixteen occasions, respectively \([i.e., d_1 = 0.2 \text{ and } d_2 = 0.8 \text{, then } TPS(ψ,d) \text{ would equal 0.54.} \]

The optimal forecast \( r_{1,K+1}^∗ \) in the two-state situation in the presence of the CSS for collections of forecasts with selected sample sizes \( K \), when the sample relative frequency \( d_1 \) and the climatological probability \( π_{1k} \) \( (k=1, \ldots, K) \), are equal and \( (b) \) unequal. The climatological probability \( π_{1k} \) \( (k=1, \ldots, K+1) \), equals 0.2 for all of the collections of forecasts and the forecaster’s judgment, \( p_{1,K+1} \), equals 0.4.

<table>
<thead>
<tr>
<th>Sample size ( K )</th>
<th>Sample relative frequency ( d_1 )</th>
<th>Optimal forecast ( r_{1,K+1}^∗ )</th>
<th>Sample relative frequency ( d_1 )</th>
<th>Optimal forecast ( r_{1,K+1}^∗ )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.20</td>
<td>0.3600</td>
<td>0.40</td>
<td>0.3700</td>
</tr>
<tr>
<td>20</td>
<td>0.20</td>
<td>0.3788</td>
<td>0.35</td>
<td>0.3862</td>
</tr>
<tr>
<td>50</td>
<td>0.20</td>
<td>0.3912</td>
<td>0.30</td>
<td>0.3936</td>
</tr>
<tr>
<td>100</td>
<td>0.20</td>
<td>0.3956</td>
<td>0.25</td>
<td>0.3962</td>
</tr>
<tr>
<td>200</td>
<td>0.20</td>
<td>0.3978</td>
<td>0.25</td>
<td>0.3981</td>
</tr>
</tbody>
</table>
climatological probability, \( p_{1k} \) \((k = 1, \cdots, K)\), are equal, as in the two previous situations, and unequal. Note that for sample sizes \( K > 50 \), the amount of hedging is less than 0.01 in both cases and that, for a given sample size, the amount of hedging is greater when the sample relative frequency and climatological probability are equal than when they are unequal. This relationship holds in \( N \)-state \((N > 2)\) as well as in two-state \((N = 2)\) situations. In addition, note, from (16), that, since the \( PS \) for the climatological probabilities, \( PS(\pi, d) \), is a maximum when the climatological probabilities, \( \pi_n \) \((n = 1, \cdots, N)\), are all equal to \( 1/N \), the amount of hedging encouraged by the CSS decreases as these probabilities approach equality (i.e., as \( \pi_1 \) approaches one-half in the two-state situation). While we have considered the hedging produced by the CSS for only a few hypothetical collections of forecasts, we believe that the hedging which is produced for these collections is representative of the hedging which would be produced for other collections with similar sample sizes.\(^{14}\)

In this regard, consider the expression for the optimal forecast \( r_{m,K+1}^* \) in (14). Note that as the sample size \( K \) increases, \( TPS(\pi, d) \) increases, and, for \( K \) large (i.e., \( K \geq 100 \)), \( TPS(\pi, d) > PS(\pi_{K+1}) \) for all \( n \) \((n = 1, \cdots, N)\). Then, \( r_{m,K+1}^* \) in (14) becomes

\[
r_{m,K+1}^* = \frac{[\rho_{m,K+1}/TPS(\pi, d)]}{\sum_{n=1}^{N} [\rho_{n,K+1}/TPS(\pi, d)]},
\]

or, since

\[
\sum_{n} \rho_{n,K+1} = 1 \quad (n = 1, \cdots, N),
\]

\[
r_{m,K+1}^* = \rho_{m,K+1}.
\]

Thus, the CSS is approximately strictly proper for large collections of forecasts.

Clearly, then, the length of the period over which a forecaster attempts to maximize his expected score in part determines the amount of hedging produced by the CSS on a particular occasion. For example, if the period is a month, during which the forecaster might make twenty forecasts \((K=20)\), then the amount of hedging would probably be greater than 0.02 during most of the period, while if the period is a season, during which the forecaster might make sixty forecasts \((K=60)\), then the amount of hedging would probably be less than 0.02 during most of the period (see Table 3). These results suggest that, if the CSS is used to evaluate precipitation probability forecasts, then the period over which the scores are accumulated should be three or more months in length.

\(^{14}\) The hedging produced by the CSS in the two-state situation is such that \( r_{1,K+1}^* < \rho_{1,K+1} \) for \( \rho_1 < 0.5 \) and \( r_{1,K+1}^* > \rho_{1,K+1} \) for \( \rho_1 > 0.5 \).

7. The standardization of scoring rules: Some questions

A number of questions arise regarding the standardization of scoring rules and we briefly consider two such questions in this section.

a. The assessment and evaluation tasks

Since assessment and evaluation can be considered to be distinct tasks and since the standardization process may often produce an improper scoring rule, could different scoring rules be used for assessing, or formulating, and for evaluating probability forecasts? For example, could the \( PS \), a strictly proper scoring rule, be used in the assessment task to discourage hedging on the part of the forecaster, while the \( SS \), an improper scoring rule, is used in the evaluation task to measure the "skill" of his forecasts? In this regard, Winkler (1969) has indicated that, while the same scoring rule need not be used in the assessment and evaluation tasks, game theoretic considerations could arise if different scoring rules were used in the two tasks. Specifically, if the forecaster is aware that he is being rewarded or penalized according to one scoring rule (e.g., the \( PS \)) to discourage hedging and evaluated according to another scoring rule (e.g., the \( SS \)) to measure "skill," then hedging may still be to his advantage rather than to his disadvantage. Thus, while different scoring rules could be used in connection with the assessment and evaluation tasks, their use might present certain problems.

b. The standardisation of strictly proper scoring rules

If we assume that the scoring rule to be used in the evaluation task is a transformation of the strictly proper scoring rule used in the assessment task, then what transformations of this scoring rule are permissible (in the sense that the transformed or standardized scoring rule is also strictly proper)? In this regard, a linear transformation of a strictly proper rule yields a strictly proper scoring rule (e.g., Winkler and Murphy, 1968). That is, if a scoring rule \( S(r, d) \) with a positive orientation is strictly proper, then the scoring rule \( S^*(r, d) \), where

\[
S^*(r, d) = aS(r, d) + b,
\]

in which \( a \) and \( b \) are constants, is strictly proper.\(^{15}\) However, since \( S^*(r, d) \) and \( S(r, d) \) are equivalent, such transformations are of little interest relative to measuring the "skill" of probability forecasts. On the other hand, a nonlinear transformation of a strictly proper scoring rule does not, in general, yield a strictly proper scoring rule. With regard to linear and nonlinear transformations of strictly proper scoring rules, the \( SS \)

\(^{15}\) Since \( S(r, d) \) has a positive orientation, \( S^*(r, d) \) has a positive orientation if \( a > 0 \) and a negative orientation if \( a < 0 \).
appears to be a linear transformation of the $PS$, in which $a = -1/PS(\pi, d)$ and $b = 1$ [see (4)]. However, $PS(\pi, d)$ is not a constant; the value of $PS(\pi, d)$ depends upon the state that obtains. That is, in general, $PS_X(\pi) \neq PS_Y(\pi)$ (j, k = 1, \ldots, N; j \neq k). Now, while $a$, in (18), must be a constant (in the sense that the value of $a$ cannot depend upon the state that obtains), no such restriction need be placed upon the constant $b$. In this regard, a modified skill score $MSS$ for the forecast $r$ can be defined as

$$MSS (r, \pi, d) = PS(\pi, d) - PS(r, d).$$  

(19)

The individual $MSS$, in (19), is a strictly proper scoring rule with a positive orientation, the range of which is the closed interval $[-2, 2]$. Note that the individual $MSS$ represents the amount by which the $PS$ for the forecast $(r_1, \ldots, r_N)$ improves upon the $PS$ for the climatological probabilities $(\pi_1, \ldots, \pi_N)$. The collective $MSS$, $CMSS$, i.e., the $MSS$ for a collection of $K$ forecasts, is then defined as

$$CMSS (r, \pi, d) = TPS(\pi, d) - TPS(r, d).$$

(20)

The $CMSS$ represents the amount by which the total $PS$ for the $K$ forecasts $r$, improves upon the $PS$ for the $K$ climatological probabilities $\pi_k$ (k = 1, \ldots, K). If we make the same assumptions in this situation as we made in connection with the $CSS$ (see Section 5), then we can easily show that the $CMSS$ is a strictly proper scoring rule with a positive orientation, the range of which is the closed interval $[-2K, 2K]$. The $MSS$ and the $CMSS$, then, appear to be appropriate scoring rules to use to measure the "skill" of probability forecasts.

Note that the $MSS$ for the forecast $r$ is simply the numerator in the expression for the $SS$ [cf. (4) and (9)]; that is,

$$SS (r, \pi, d) = MSS (r, \pi, d)/PS(\pi, d).$$

(21)

Now, the $SS$ appears to be a linear transformation of the $MSS$, in which $a = 1/PS(\pi, d)$ and $b = 0$. However, since $PS(\pi, d)$ is not a constant, the $SS$ is an improper scoring rule. A similar relationship exists among the $CSS$, the $CMSS$, and the $TPS$ [cf. (7) and (20)]. In this case, although the $CMSS$ is strictly proper, the $CSS$ is improper because $TPS(\pi, d)$ is not a constant. Thus, only certain nonlinear transformations of strictly proper scoring rules are permissible; for example, those nonlinear transformations of the form $S^*(r, d) = aS(r, d) + b$, in which $a$ is a constant and $b$ depends upon the climatological probabilities.

8. Summary and conclusion

In this paper we have examined the individual and collective skill scores, $SS$ and $CSS$, respectively, both of which are standardized versions of the Brier, or probability, score, $PS$, and have been used to measure the "skill" of probability forecasts. The $SS$ is defined in terms of individual forecasts, while the $CSS$ is defined in terms of collections of forecasts. We have shown that the $SS$ and the $CSS$ are, in general, improper scoring rules, and that, as a result, both the $SS$ and the $CSS$ encourage hedging on the part of forecasters.

We have also presented the results of a preliminary investigation of the nature of the hedging produced by the $SS$ and the $CSS$. In this regard, in terms of both probability and expected score, the $SS$ may encourage a considerable amount of hedging, particularly when the climatological probabilities are small (or large). However, the $SS$ appears to be slightly less likely to encourage hedging than the $PS$ is to discourage hedging. The $CSS$, on the other hand, encourages only a modest amount of hedging, and even this hedging decreases as the sample size of the collection of forecasts increases. In fact, we have shown that the $CSS$ is approximately strictly proper for large collections of forecasts.

Finally, we have briefly considered two questions related to the standardization of scoring rules: 1) the use of different scoring rules in the assessment and evaluation tasks, and 2) the transformation of strictly proper scoring rules. First, while different scoring rules can be used in the assessment and evaluation tasks, e.g., the $PS$ in the assessment task to discourage hedging and the $SS$ in the evaluation task to measure "skill," the use of different scoring rules in these two tasks may introduce game theoretic considerations which could produce hedging on the part of forecasters even in the presence of the $PS$. Second, while a linear transformation of a strictly proper scoring rule is, of course, strictly proper, certain nonlinear transformations may also be strictly proper. Specifically, the additive constant in the linear transformation can be a function of the climatological probabilities. In this regard, the modified skill scores, $MSS$ and $CMSS$, which represent the difference between the $PS$ for the relevant climatological probabilities and the $PS$ for the forecast or forecasts of concern and which, as a result, are strictly proper, appear to be appropriate scoring rules to use to measure the "skill" of probability forecasts.

Acknowledgments. The author would like to acknowledge the valuable comments of Lawrence A. Hughes and Robert L. Winkler on earlier versions of this manuscript.

REFERENCES


---

\[16\] MSS(\pi, d) has been described by Sanders (1958, 1963).


