Proper Scores for Probability Forecasts Can Never Be Equitable

IAN T. JOLLIFFE AND DAVID B. STEPHENSON

School of Engineering, Computing, and Mathematics, University of Exeter, Exeter, United Kingdom

(Manuscript received 8 March 2007, in final form 31 July 2007)

ABSTRACT

Verification is an important part of any forecasting system. It is usually achieved by computing the value of some measure or score that indicates how good the forecasts are. Many possible verification measures have been proposed, and to choose between them a number of desirable properties have been defined. For probability forecasts of a binary event, two of the best known of these properties are propriety and equitability. A proof that the two properties are incompatible for a wide class of verification measures is given in this paper, after briefly reviewing the two properties and some recent attempts to improve properties for the well-known Brier skill score.

1. Introduction

Forecast verification is a crucial aspect of any prediction system. It is important to assess the quality of forecasts if improvements are to be made. A large number of verification measures have been suggested (Jolliffe and Stephenson 2003). To narrow the range of possible measures, a number of desirable properties of measures have been proposed and generally accepted (Murphy 1993; Mason 2003). For probability forecasts of a binary event, the two most frequently cited properties are *propriety* (Winkler and Murphy 1968) and *equitability* (Gandin and Murphy 1992). A score is *proper* if it is optimized only when a forecaster predicts according to his/her true beliefs and is *equitable* if the same expected score is achieved for all unskilled forecasts of a certain type—for example, constant forecasts.

The widely used Brier score (Brier 1950) is proper but not equitable (Mason 2004), and recent research has attempted to adapt the score to circumvent this "nonequitability" (Mason 2004; Müller et al. 2005; Weigel et al. 2007). It does not seem to have been noted in previous studies that propriety and equitability are incompatible. A proof that it is not possible for a verification score to simultaneously possess both properties is given in section 4 of this paper. Before that, Brier score, propriety, and equitability are defined in section 2, and recent attempts to alleviate or remove the nonequitability of the Brier skill score are briefly discussed in section 3.

2. Definitions

a. Probability forecasts of binary events

Suppose that a set of probability forecasts is made of n binary events of interest such as "precipitation tomorrow" or "damaging frost next month." Denote the n forecasts by $\{p_1, p_2, \ldots, p_n\}$, where each p_i is a probability between 0 and 1. The corresponding observations $\{x_1, x_2, \ldots, x_n\}$ are coded as 1 if the event occurs and 0 if it does not. To assess the quality of the forecasts, various measures or scores can be constructed that quantify the difference between the set of forecasts and the corresponding set of observations.

The best known of these scores, and one of the simplest, is the Brier score (Brier 1950), which is based on the sum of squared differences between the forecasts and observations:

$$B = \frac{1}{n}\sum_{i=1}^{n}(p_i - x_i)^2. \quad (1)$$

There are many other possibilities, such as the linear score or logarithmic score, defined in Table 1. To choose between the wide variety of possible scores, some desirable properties have been proposed. This idea of assessing the verification measures themselves is

Corresponding author address: Prof. Ian T. Jolliffe, 30 Woodvale Road, Gurnard, Cowes, Isle of Wight, PO31 8EG, United Kingdom.
E-mail: i.t.jolliffe@exeter.ac.uk

DOI: 10.1175/2007MWR2194.1

TABLE 1. Three scores for probability forecasts and associated functions of the scores.

Score	Definition	$S_0(p)$	$S_1(p)$	S_0'/S_1'
Linear	$\frac{1}{n}\sum_{i=1}^{n} \lvert x_i - p_i \rvert$	p	$(1 - p)$	-1
Brier	$\frac{1}{n}\sum_{i=1}^{n} (x_i - p_i)^2$	p^2	$(1 - p)^2$	$\frac{-p}{(1 - p)}$
Logarithmic	$-\frac{1}{n}\sum_{i=1}^{n} \log(\lvert x_i + p_i - 1 \rvert)$	$-\log(1 - p)$	$-\log(p)$	$\frac{-p}{(1 - p)}$

sometimes known as *metaverification* (Murphy 1996). Two such properties that are particularly relevant to probability forecasts are propriety and equitability.

b. Proper scores and hedging

The concept of proper scores dates back at least as far as Winkler and Murphy (1968). Suppose that, as above, a forecaster makes forecasts $\{p_1, p_2, \ldots, p_n\}$. On these n forecast occasions, the forecaster's true beliefs of the probability of the event are $\{q_1, q_2, \ldots, q_n\}$. If any q_i is different from p_i, then the forecaster is said to be *hedging* the forecast (Murphy and Epstein 1967). Hedging is beneficial to the forecaster if using it improves the expected (long-term average) value of the score that is being used to assess his or her forecasts. A score is proper if the forecaster's expected value for the score is optimized by taking $p_i = q_i$. In other words, the forecaster cannot benefit from hedging, that is, by making a forecast other than his or her true belief. If the expected value is *uniquely* optimized by $p_i = q_i$, the score is *strictly proper*. It is generally accepted that propriety (a score being proper) is highly desirable—it is undesirable for a score to allow forecasters to benefit from hedging their forecasts.

The topic of proper scores has been discussed in some detail in both the meteorological and statistical literature (see, e.g., Winkler and Murphy 1968; Savage 1971; Gneiting and Raftery 2007). The latter authors, among others, discuss the characterization of proper scoring rules. There exists an infinite number of such rules, but only a few have been explicitly described. The Brier score is by far the best known of these, but the logarithmic score in Table 1 is also proper.

Each of these scores is often represented as a skill score, which measures the skill of forecasts relative to a baseline or reference forecasting system. Transforming a score into a skill score often loses propriety. For example, the Brier skill score may be written:

$$B_s = 1 - \frac{B}{B_{\text{ref}}}, \tag{2}$$

where B_{ref} is the value of B for some unskilled reference forecasting system; B_s has the property that it takes the value 0 for the reference forecasting system and is 1 for perfect forecasts. Reference forecasts can be chosen in a number of ways (Jolliffe and Stephenson 2003). One possibility is to issue forecasts according to a random mechanism that is unrelated to any previous observations. Such forecasts are clearly unskillful. A second possibility is to use persistence, for which the forecast for the next time period is identical to what is currently observed. Strictly this is not unskillful in most situations but it is represents a simple strategy that any forecasting system should improve upon. The most common reference forecast is "climatology," in which the forecast is always the long-term probability of the event of interest, θ, sometimes known as the base rate. The B_{ref} for climatology is often taken as $\theta(1 - \theta)$, the value of B achieved in the long run if θ is always forecast (Toth et al. 2003). If θ is based on data different from those used to assess the forecasts, then B_{ref} is a constant, not depending on the forecasts or observations. Hence B_s shares the properties of B and is proper. In practice θ is usually estimated by "sample climatology," $\bar{x}$, the proportion of times that forecasted event occurs in the sample and, in this case, the score is no longer proper (Murphy 1973), except asymptotically, though Mason (2004) erroneously claims that it is.

c. Equitable scores

Another widely known desirable property is equitability. This was first defined by Gandin and Murphy (1992), though the concept was used many years earlier in devising a verification measure for forecasts of categorical events (Jolliffe and Foord 1975). It is based on the idea that all unskillful forecasts of a certain type should have the same expected score. Specifically, constant forecasts that are always the same ($p_i = p$, $i = 1, 2, \ldots, n$) are clearly unskillful, as are forecasts that randomly choose a forecast from some distribution on the interval [0, 1] (constant forecasts are a special case in which all the probability in the interval is concen-

trated at a single point). It would seem odd if two different forecasting strategies from this class have different expected scores, as all such strategies are "equally unskillful." Hence equitability is considered desirable.

3. Nonequitability of the Brier skill score

None of the proper scores that have been suggested in the literature are equitable (see section 4). In particular, with climatology as a reference forecast, B_s can be negative for other unskillful (potential reference) forecasts. This is undesirable in its own right, but from a practical point of view it also means that forecasting systems with skill may have values of B_s close to zero and hence look distinctly unimpressive. This is referred to in the literature as a negative bias in B_s. Both Mason (2004) and Müller et al. (2005) have suggested modifications of B_s to alleviate or circumvent such negative bias.

Müller et al. (2005) discuss the situation in which an ensemble of m forecasts is generated and the probability forecast for the event of interest is the proportion of ensemble members for which the event occurs. They consider the ranked probability skill score (RPSS), which extends the Brier skill score to more than two categories. To reduce the negative bias in this skill score, Müller et al. (2005) replace the usual reference score by one based on resampling of the climatology. This turns out to be equivalent to using a reference forecasting strategy in which a binomial random variable is generated with m trials and probability of success θ, where m is the number of ensemble members and θ the climatological probability or base rate (Weigel et al. 2007; see also Ferro et al. 2008). This alleviates the problem of negative bias of the RPSS (and, as a special case, the Brier skill score) for ensemble forecasts although as will be reported elsewhere, there may be advantages in slightly adjusting the probability of success in the reference forecast away from θ.

Mason (2004) proposes a different approach. One of his objectives is to ensure that all unskillful forecasts have a nonnegative expected score. This addresses the problem of negative bias, though it now has the opposite problem that positive values of the score do not necessarily imply skill. His reference forecast is different from those proposed previously in that it depends on the forecasts themselves, so the reference forecast changes as the set of forecasts being assessed changes. By allowing this dependence on forecasts, Mason (2004) derives a variant of the Brier skill score that he claims is equitable.

4. A no-go theorem for propriety and equitability

Many scores are additive and so can be written as the mean "loss" over all n forecast times:

$$S = \frac{1}{n}\sum_{i=1}^{n} S(p_i, x_i).$$

For example, the Brier score is additive and can be written in this way with loss $S(p, x) = (p - x)^2$. For each forecast time, the loss can either be $S_1(p) = S(p, 1)$ when the observed event occurs, or $S_0(p) = S(p, 0)$ when it does not. The two functions $S_0(p)$ and $S_1(p)$ fully define the loss: $S(p, x) = (1 - x)S_0(p) + xS_1(p)$. Examples of $S_0(p)$ and $S_1(p)$ for three scores are given in Table 1.

Most scores are additive since one generally assumes that the two loss functions at any particular time do not depend on the values of x and p at the other times. However, this is not the case for skill scores in which the reference forecast depends on the other values of either x or p. For example, the Brier skill score based on climatological mean forecasts has a denominator that depends on the sample variance $\bar{x}(1 - \bar{x})$ of all the x values and so depends on x in both the numerator and denominator; it cannot therefore be written as a mean loss over all previous forecast times. In the asymptotic limit of large sample size, the denominator in the skill score tends to a constant, so the skill score then tends to an additive sum. This leads to the interesting result that, although the Brier score is proper, the Brier skill score is only proper in the asymptotic limit of large sample size. In what follows, we will assume that a score can be written additively, although this is generally only true in the asymptotic limit for *skill* scores.

If the forecaster believes the probability $\Pr(x = 1)$ of a future event at a certain time is q, which is strictly between 0 and 1, then the expected value of the score when p is forecast is

$$S(p, q) = (1 - q)S_0(p) + qS_1(p), \tag{3}$$

where expectation here is with respect to the forecaster's belief. Probability q is a subjective probability in that it reflects the belief of a forecaster and so can be different for different forecasters. It is sometimes misleadingly referred to as the *true probability* of the event; however, it is not unique and is only *true* for a particular forecaster for an event at a certain time.

For propriety, we require that this expected score is minimized when $p = q$, in other words, when the forecaster issues a forecast that matches his or her belief. For fixed q, either the expected score is a monotonic function of p between 0 and 1, in which case the score cannot be proper because hedging to 0 or 1 is beneficial,

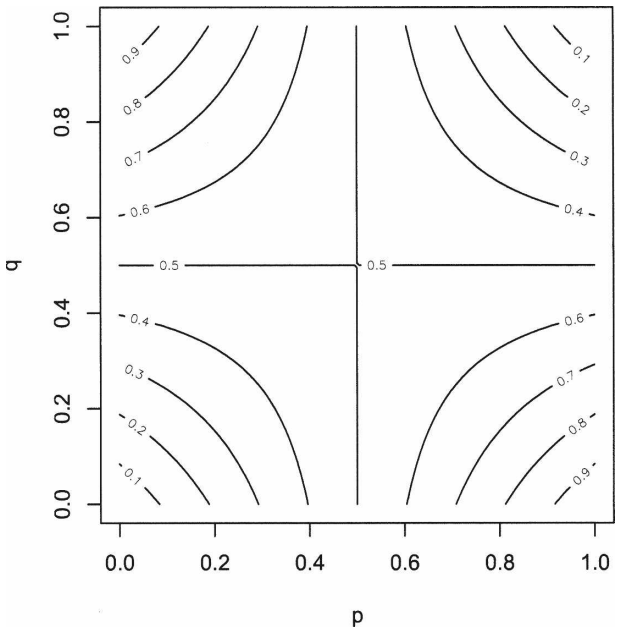

FIG. 1. Contour plot of $S(p, q)$ in Eq. (3) for the linear score.

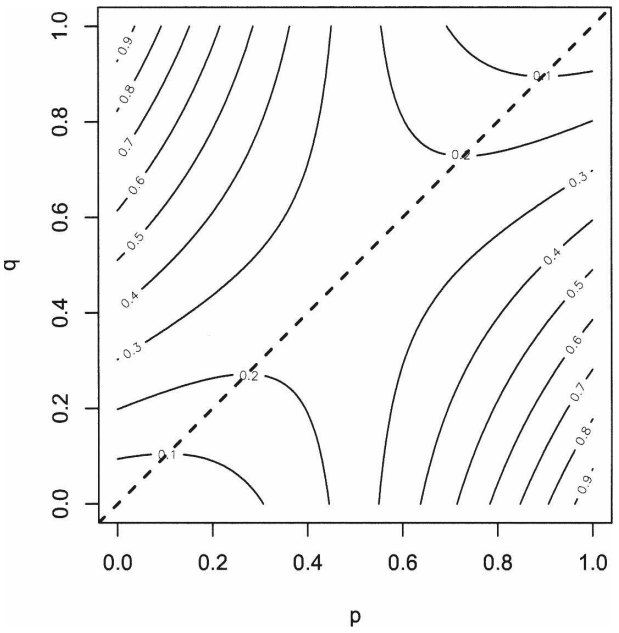

FIG. 2. As in Fig. 1, but for the Brier score.

or it has a minimum strictly between 0 and 1. In the latter case, the derivative of the expected score in Eq. (3) with respect to p must be 0 at the minimum. Hence,

$$\frac{\partial S}{\partial p} = (1 - q)\frac{\partial S_0}{\partial p} + q\frac{\partial S_1}{\partial p} = 0, \tag{4}$$

so, for propriety:

$$\left.\frac{S_0'}{S_1'}\right|_{p=q} = \frac{-q}{(1-q)} = \frac{-p}{(1-p)}, \tag{5}$$

where the prime denotes a derivative with respect to p. For strict propriety the ratio of derivatives on the left is equal to the ratio on the right *only* when $p = q$.

Now consider equitability. For a score to be equitable, any forecast of the form "always issue the same forecast p" should have the same climatological time-mean score for any p. The expected score is

$$S(p, \theta) = (1 - \theta)S_0(p) + \theta S_1(p), \tag{6}$$

where θ is the base rate, and expectation is now with respect to climatology (i.e., the average score in the long run if p is always forecast). Hence, for equitability, the derivative of $S(p, \theta)$ must be 0 for all p between 0 and 1. Therefore,

$$\frac{S_0'}{S_1'} = \frac{-\theta}{(1-\theta)} \tag{7}$$

for all p between 0 and 1, where θ is a single fixed value. Equations (5) and (7) are clearly incompatible. Equation (7) states that the ratio of derivatives is constant for all possible forecasts p, whereas Eq. (5) says that the ratio varies as p varies. Hence, it is impossible to achieve both equitability and propriety.

Examples

Table 1 and Figs. 1–3 illustrate the ideas of propriety and equitability for three scores: a linear score, a logarithmic score, and the Brier (quadratic) score.

The first column of the table presents the definition of each score, given a set of n forecasts $\{p_1, p_2, \ldots, p_n\}$

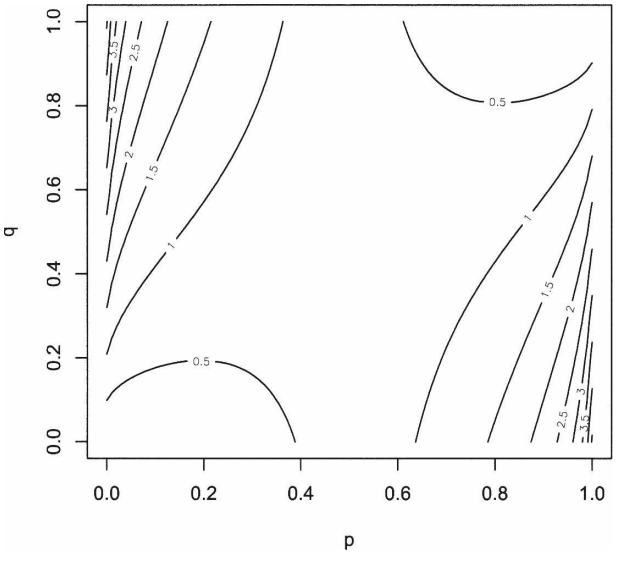

FIG. 3. As in Fig. 1, but for the logarithmic score.

and corresponding observations $\{x_1, x_2, \ldots, x_n\}$. The next three columns give the form of $S_0(p)$, $S_1(p)$, and the ratio of derivatives S_0'/S_1' for each measure. The figures show the form of $S(p, q)$ for the three scores and illustrate geometrically the necessary conditions for equitability and propriety, that is, where $S(p, q) = S(p, \theta)$.

Figure 1 displays $S(p, q)$ for the linear score, for which $S_0(p) = p$ and $S_1(p) = (1 - p)$. Here the function is constant for all p when $q = \theta = ½$. Hence, for this value of the base rate, but for no other, the linear score is equitable. For any value of q, the minimum value of $S(p, q)$ is achieved when p is zero or unity, depending on whether q is less than or greater than ½. Thus it is always advantageous to hedge forecasts to 0 or 1, and the score is not proper.

Figure 2 shows $S(p, q)$ for the Brier score, for which $S_0(p) = p^2$ and $S_1(p) = (1 - p)^2$. Here there is no value of θ for which $S(p, \theta)$ is constant for all p, so the score is never equitable. However, the minimum value of $S(p, q)$ is achieved on the 45° line in Fig. 2, and only on that line for every value of q, so that the score is strictly proper. Although Fig. 3, for the logarithmic score with $S_0(p) = -\log(1 - p)$, $S_1(p) = -\log(p)$, looks less simple than Fig. 2, its underlying structure is the same, illustrating propriety and nonequitability.

For both Figs. 2 and 3, if the figures are considered as contours on a topographical map, then walking from west to east is uphill until the 45° line and downhill thereafter, demonstrating propriety. There is no west–east traverse that is flat, so equitability is impossible. However, in Fig. 1, the west–east traverse is flat for $q = 0.5$, demonstrating equitability in this case. For any other value of q, the traverse will be either all uphill (for $q < 0.5$) or all downhill ($q > 0.5$), ruling out propriety.

At this point, it should be noted again that our no-go theorem is restricted to a certain class of verification measures, namely those where the measure for a set of n forecasts can be written as the sum or average of scores for each individual forecast. Using the notation of section 4, the overall score S can be written as

$$S = \frac{1}{n} \sum_{i=1}^{n} S_{\delta_i}(p_i), \tag{8}$$

where p_i is the ith probability forecast and $\delta_i = 1$ if the event occurs for the ith forecast and equals 0 otherwise. Furthermore, the score for any individual forecast must not depend on any of the other $(n - 1)$ forecasts in the set or their corresponding observations.

Many well-known verification measures fall into this class, as demonstrated by Table 1, but their skill score versions often do not. This is because, as discussed above, the transformation used to convert a measure into a skill score often involves all the data, as when sample climatology is used as a reference forecast.

Our intuition is that propriety and equitability are incompatible for all verification measures, including such skill scores, and this view is reinforced by the work described in section 3. However, demonstration of the wider result awaits further research.

5. Discussion

It would be ideal to have a verification score that is both proper and equitable, but the previous section shows, for a large class of measures, that this is impossible for probability forecasts of a binary event. Both properties have their appeal, but our view is that propriety is the more fundamental requirement. The implication is that equitability is a largely irrelevant property for probabilistic forecasts. However, it should not be forgotten that, unlike propriety, equitability can also be defined for deterministic forecasts and is highly relevant for such forecasts. Indeed, the idea of equitability was first introduced for deterministic categorical forecasts (Gandin and Murphy 1992) and later, perhaps mistakenly, adopted for probabilistic forecasts.

The loss of equitability for probability forecasts has important implications for how one interprets whether or not a forecasting system has skill. Nonequitability means that different unskillful forecasts can give different scores and hence there is no unique absolute benchmark against which to measure skill. For example, the skill of a forecasting system may have negative skill when compared with a constant probability climatological forecast and yet have positive skill when compared with random probability forecasts (Mason 2004; Müller et al. 2005; Weigel et al. 2007). How should one then decide whether the forecasting system really has skill? Perhaps the most rational approach is to demand that the score is better than the best score of all possible unskillful forecasts. To be able to do this, one needs to think carefully about how best to optimize the score using unskillful forecasts. For scores where the loss functions $S_0(p)$ and $S_1(p)$ are convex functions of p, the best unskillful forecasts are those that issue constant probability values, albeit not necessarily the base rate. Alternatively, one can avoid some of these difficulties by eschewing any mention of the word "skill" and simply present scores for different forecasting systems: it is still possible to order forecasting systems based on pairwise comparison of scores.

Acknowledgments. We are grateful to an anonymous reviewer, whose comments led us to substantially re-

think some of our views. Much of the work leading to this paper was done while ITJ was funded by a NERC Environmental Mathematics and Statistics Discipline Bridging Award (NER/T/S/2003/00126).

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteor. Appl.,* in press.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.,* **120,** 361–370.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.,* **102,** 359–378.

Jolliffe, I. T., and J. F. Foord, 1975: Assessment of long-range forecasts. *Weather,* **30,** 172–181.

——, and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 254 pp.

Mason, I. B., 2003: Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 37–76.

Mason, S. J., 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **132,** 1891–1895.

Müller, W. A., C. Appenzeller, F. J. Doblas-Reyes, and M. A. Liniger, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Climate,* **18,** 1513–1523.

Murphy, A. H., 1973: Hedging and skill scores for probability forecasts. *J. Appl. Meteor.,* **12,** 215–223.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

——, and E. S. Epstein, 1967: A note on probability forecasts and "hedging." *J. Appl. Meteor.,* **6,** 1002–1004.

Savage, L. J., 1971: Elicitation of personal probabilities and expectations. *J. Amer. Stat. Assoc.,* **66,** 783–801.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 137–163.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **135,** 118–124.

Winkler, R. L., and A. H. Murphy, 1968: "Good" probability assessors. *J. Appl. Meteor.,* **7,** 751–758.