A Note on the Attributes of Probabilistic Predictions and the Probability Score[1,2]

EDWARD S. EPSTEIN AND ALLAN H. MURPHY

University of Michigan, Ann Arbor

4 November 1964 and 14 January 1965

The "value"[3] of an evaluation measure for probabilistic predictions depends upon the ability of the measure to ascertain the extent to which the predictions possess certain "desirable"[3] attributes. Since the form of the measure depends upon the particular attributes selected for consideration, the attributes should be identified and defined prior to the construction of a measure. Bross (1953) identified and defined the attributes *validity* and *sharpness*,[4] in terms of which measures can be constructed. Evaluation measures for probabilistic predictions have been constructed by Brier, Holloway and Woodbury, and Miller among others. These measures, with the possible exception of the *validity measure* of Miller (1962), have not been constructed with particular reference to selected attributes of the probabilities. Still, the measures do appear to provide a means of examining the validity and/or sharpness of predictions (Miller, 1962). In particular, the *probability score* (Brier, 1950) appears to provide a means to evaluate validity *and* sharpness (Miller, 1962; Sanders, 1963). However, different interpretations of the terms validity and sharpness as well as of the probability score have complicated the evaluation of the latter as a measure (Murphy and Epstein, 1964). The purposes of this note are to propose precise definitions for two attributes of probabilistic predictions and for a natural measure for each attribute; and to indicate the relationship between the measures, as defined, and the probability score. The geometrical framework within which the definitions are developed is the "probability triangle" proposed by de Finetti (1962).

Consider an equilateral triangle $O_1O_2O_3$ (Fig. 1). A point P within or on the triangle may be (uniquely) identified by the set $\{p_i;\ i=1, 2, 3\}$ of numbers, where p_i represents the distance from P to the side O_jO_k ($j,\ k=1, 2, 3;\ j,\ k\neq i$). The sum of the distances is invariant for all points within or on the triangle, and is, in fact, equal to the altitude h of the triangle, *i.e.*,

$$\sum_{i=1}^{3} p_i = h.$$

If h equals one,

$$\sum_{i=1}^{3} p_i = 1,$$

and

$$0 \leq p_i \leq 1,\ i=1, 2, 3.$$

Suppose the vertex O_i of the triangle is associated with the occurrence of an event E_i, where the set $\{E_i;\ i=1, 2, 3\}$ represents a set of mutually exclusive and exhaustive events. Then, the set $\{p_i;\ i=1, 2, 3\}$ of distances identifying the point P defines a probability statement on the set $\{E_i;\ i=1, 2, 3\}$ of events where the distance p_i represents the probability of occurrence of the event E_i.

If $D_i (i=1, 2, 3)$ is the distance from the point P to the vertex O_i, then

$$D_i = \frac{2\sqrt{3}}{3}(p_j^2 + p_k^2 + p_j p_k)^{\frac{1}{2}},\quad j,k \neq i,\quad j \neq k.$$

De Finetti (1962) has proposed that a *score* or *penalty*, $-S$, be associated with the probability statement P when the event E_i obtains, where

$$S = \tfrac{3}{4}D_i^2 = (p_j^2 + p_k^2 + p_j p_k). \tag{1}$$

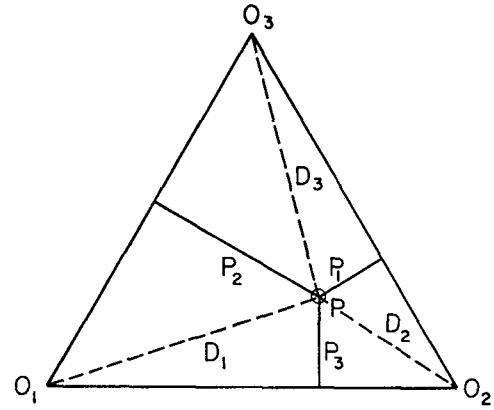

FIG. 1. The probability triangle.

[1] This research was supported in part by the United States Weather Bureau under Contract Cwb-10847.

[2] Publication Number 81 from the Department of Meteorology and Oceanography, The University of Michigan.

[3] The terms *value* and *desirable* remain undefined since the identification and definition of the *purposes* for evaluation are not to be considered. Such terms may be defined only with reference to a *particular* purpose.

[4] Validity refers to the resemblance, on an independent collection of predictions, between the probabilities assigned to, and the observed relative frequencies of, the events, while sharpness refers to the discrimination among events exhibited by the probabilities.

The values of S range from zero, when the probability statement assigns probability one to the event E_i, to one, when the probability statement assigns probability one to either of the other events. The vertex O_i is associated with the event E_i because the probability statement (*i.e.*, the set of distances) $\{p_i=1,\ p_j=0,\ p_k=0\}$ places the point P at the vertex O_i.

Attributes and measures will now be defined within this geometrical framework. Consider N points (within or on the equilateral triangle) to represent a collection of N probabilistic predictions. The collection of N points will have a centroid $\bar{P}$ defined by the set $\{\bar{p}_i;\ i=1, 2, 3\}$ of distances, where

$$\bar{p}_i=\frac{1}{N}\sum_{n=1}^{N} p_{in},\quad i=1, 2, 3.$$

Let $N_i(i=1, 2, 3)$ represent the number of occasions in N that the event E_i obtains. Then,

$$\sum_{i=1}^{3} N_i=N.$$

If the relative frequency of occurrence of the event E_i is f_i, then

$$f_i=\frac{N_i}{N},\quad i=1, 2, 3,$$

and

$$\sum_{i=1}^{3} f_i=1.$$

The point F in the triangle defined by the set $\{f_i;\ i=1, 2, 3\}$ of distances represents the climatological probability statement for the collection. The *bias*[5] B of the predictions may be defined in terms of the distance between the point P and the point F, *i.e.*,

$$B=[(p_i-f_i)^2+(p_j-f_j)^2+(p_i-f_i)(p_j-f_j)]^{\frac{1}{2}},\quad i, j=1, 2, 3;\ i\neq j. \quad (2)$$

The values of B range from zero, for a completely unbiased collection of predictions, to one, for a completely biased collection of predictions.

The collection of predictions may be divided into M sub-collections by defining conditions on the values assumed by the members of the set $\{p_i;\ i=1, 2, 3\}$. In terms of the geometrical framework such a procedure subdivides the region defined by the probability triangle. A natural subdivision of the region is that in which $M=9$ subregions are defined by the conditions

[5] The term bias refers, in general, to the difference between the expected value of a statistic and the true value of the parameter of which it is an estimator. Thus, the attribute *bias* is properly the *difference* between P and F. The Euclidean distance between the points P and F (Eq. (2)) is then a natural measure of this attribute. Brier (1950) and Ackoff (1962) have previously proposed the term bias as an attribute of predictions, but neither suggested this particular measure.

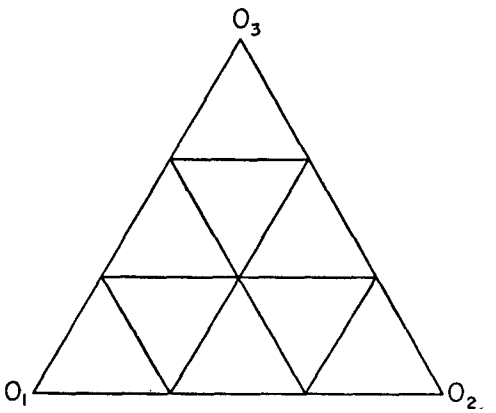

FIG. 2. A possible subdivision of the region defined by the probability triangle.

$0\leq p_i\leq\frac{1}{3}$, $\frac{1}{3}<p_i\leq\frac{2}{3}$, $\frac{2}{3}<p_i\leq 1$; $i=1, 2, 3$ (Fig. 2). A value of the bias $B_m(m=1, \ldots, M)$ may be computed (utilizing (2) with the appropriate points $\bar{P}_m$ and F_m) for each prediction sub-collection. The M values of the bias will indicate which sub-collections contain relatively biased or relatively unbiased predictions. Further, the sum of the M biases provides a measure of the total bias of the collection of predictions. (Note that the value of the bias B, Eq. (2), is the length of the weighted vector sum of the line segments whose lengths are B_m.)

A second attribute of probabilistic predictions, *validity*, may now be defined as the *degree of accuracy* of a prediction, i.e., the *difference* between the point P defining the probability statement (prediction) and the vertex O_i associated with the event E_i which subsequently obtains. Within this geometrical framework, a natural measure of this attribute of the predictions is $(S)^{\frac{1}{2}}$, Eq. (1). The overall validity V of a collection of N predictions is then

$$V=\frac{1}{N}\sum_{n=1}^{N}(S_n)^{\frac{1}{2}}.$$

The values of V range from zero, for a completely valid collection of predictions, to one, for a completely invalid collection of predictions.

The probability score PS (Brier, 1950) for a prediction defined on the set $\{E_i;\ i=1, 2, 3\}$ of mutually exclusive and exhaustive events is

$$PS=\sum_{i=1}^{3}(p_i-O_i)^2,$$

where p_i is the probability assigned to event E_i and O_i is one if event E_i obtains and zero otherwise. If event E_i obtains the probability score becomes

$$\begin{aligned} PS&=(p_i-1)^2+p_j^2+p_k^2,\\ &=(-p_j-p_k)^2+p_j^2+p_k^2,\\ &=2(p_j^2+p_k^2+p_jp_k). \end{aligned}$$

TABLE 1. Prediction collections.

Prediction number	Probability of event number 1	2	3	Observed event number	Probability score PS	Bias B	Validity V
				Collection I			
1	1/3	1/3	1/3	1	2/3		$\sqrt{3}/3$
2	1/3	1/3	1/3	2	2/3		$\sqrt{3}/3$
3	1/3	1/3	1/3	3	2/3		$\sqrt{3}/3$
Average	1/3	1/3	1/3		2/3	0(0.000)	$\sqrt{3}/3$(0.577)
				Collection II			
1	1/3	1/3	1/3	1	2/3		$\sqrt{3}/3$
2	1/3	1/3	1/3	1	2/3		$\sqrt{3}/3$
3	1/3	1/3	1/3	1	2/3		$\sqrt{3}/3$
Average	1/3	1/3	1/3		2/3	$\sqrt{3}/3$(0.577)	$\sqrt{3}/3$(0.577)
				Collection III			
1	1	0	0	1	0		0
2	0	1	0	1	2		1
3	0	0	1	1	2		1
Average	1/3	1/3	1/3		4/3	$\sqrt{3}/3$(0.577)	2/3(0.667)

Thus,

$$PS=2S,$$

so that the probability score and de Finetti's score are essentially equivalent. The probability score is then a measure of the distance from the point representing a prediction to the (*a posteriori*) "desirable" vertex and, as such, is a measure of the attribute validity (as defined above).

The proposed measures are illustrated in terms of an evaluation of three small collections of predictions (Table 1). Observe that the predictions in collections I and II are equally valid. However, the predictions in collection I are completely unbiased while the predictions in collection II are biased. On the other hand, the predictions in collections II and III are equally biased, but collection II contains more valid predictions. Note that the predictions in collection III are perfectly *sharp* (footnote 4), while the predictions in collections I and II exhibit a complete absence of discrimination among the events. The authors have chosen neither to concern themselves with nor to attempt to measure this attribute. A comparison of prediction collections II and III also illustrates the difference between an average of a distance (V) and the average of the square of a distance ($\overline{PS}$). The ratio of the validities for collections II and III is 0.87, while the ratio of the probability scores is 0.50.

The geometry and the measures defined above have been restricted to two dimensions, but may be extended directly to higher dimensions. The appropriate geometric figure in three dimensions is a regular tetrahedron (four mutually exclusive and exhaustive events). In general, the appropriate figure in $K-1$ dimensions is a regular polytope (K mutually exclusive and exhaustive events), although for $K>4$ the geometrical framework is difficult to perceive. Nevertheless, one can show analytically that the proper Euclidean distances, in K-1 dimensions, when the event $E_l(l=1, \ldots, K)$ obtains, are given by

$$S=\sum_{\substack{i=1\\ i,\,j\neq l}}^{K}\sum_{j=1}^{K} p_i p_j$$

and

$$B^2=\sum_{i=1}^{K-1}\sum_{j=1}^{K-1}(\bar{p}_i-f_i)(\bar{p}_j-f_j),$$

and that

$$S=\tfrac{1}{2}PS=\tfrac{1}{2}\sum_{i=1}^{K}(p_i-O_i)^2.$$

REFERENCES

Ackoff, R. L., 1962: *Scientific Method: Optimizing Applied Research Decisions*. New York, John Wiley and Sons, 393–394.

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

Bross, I. D. J., 1953: *Design for Decision*. New York, The Macmillian Co., 46–53.

De Finetti, B., 1962: Does it make sense to speak of 'good probability appraisers'? *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, New York, Basic Books, Inc., 357–364.

Miller, R. G., 1962: Statistical prediction by discriminant analysis. *Meteor. Monogr.*, **4**, No. 25, 9–10.

Murphy, A. H., and E. S. Epstein, 1965: The evaluation of probability prediction systems. *J. Appl. Meteor.*, in press.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.