Second-Order Probabilities and Strictly Proper Scoring Rules[1]

ALLAN H. MURPHY[2]

Dept. of Meteorology and Oceanography, University of Michigan, Ann Arbor

AND CARL-AXEL S. STAËL VON HOLSTEIN[3]

Dept. of Engineering-Economic Systems, Stanford University, Stanford, Calif.

(Manuscript received 1 September 1970, in revised form 10 November 1970)

ABSTRACT

Some forecasters apparently subscribe to a model of the subjective probability forecasting process in which their judgments are expressed in terms of "second-order" probabilities. First, we briefly consider the nature of these second-order probabilities and describe the second-order model, and then we demonstrate that strictly proper scoring rules encourage forecasters who subscribe to the second-order model to make their forecasts correspond to their *expected* judgments.

1. Introduction

In the *standard* model of the subjective probability forecasting process, a forecaster expresses his judgment concerning (say) the occurrence of (measurable) precipitation in terms of a number p ($0 \leqslant p \leqslant 1$). Further, this model prescribes that the forecaster modifies p when he receives additional relevant information. However, recent discussions with National Weather Service forecasters[4] have indicated that this model may not provide an adequate description of the behavior of certain forecasters. Apparently, these forecasters express their judgments in terms of a probability distribution $F(p)$ on the probability p, and modify $F(p)$ upon the receipt of additional relevant information. Since p is itself a probability, the distribution $F(p)$ consists of "second-order" probabilities. Thus, we can refer to the standard model as a *first-order* model and to the model which involves second-order probabilities as a *second-order* model.

Strictly proper scoring rules have been formulated as a means of encouraging a forecaster who subscribes to the first-order model to make his forecast correspond to his judgment. Specifically, in the presence of a strictly proper scoring rule, such a forecaster can maximize his expected score[5] only if he makes his (precipitation probability) forecast r correspond exactly to his judgment p. The purpose of this paper is to investigate the effect of strictly proper scoring rules on the behavior of forecasters who subscribe to the second-order model.

We briefly consider the nature of second-order probabilities and describe the second-order model in Section 2. In Section 3 we investigate the effect of strictly proper scoring rules on the behavior of forecasters who subscribe to the second-order model for both a special and the general situation. Section 4 consists of a brief summary and conclusion.

2. Second-order probabilities and the second-order model

As indicated in Section 1, the distribution $F(p)$ consists of second-order probabilities. These probabilities can be thought of as expressing, in quantitative, probabilistic terms, the forecaster's "uncertainty" over the possible judgments concerning the occurrence of precipitation. For example, a particular forecaster on a particular occasion may believe that "the probability that the precipitation probability p is 0.2 is greater than the probability that p is 0.5." Savage (1954, pp. 57–60) raises certain questions of a theoretical nature related to the introduction of second-order probabilities as a means of expressing this uncertainty. How-

[1] Contribution No. 186 from the Department of Meteorology and Oceanography, University of Michigan.

[2] Supported in part by the National Science Foundation (Atmospheric Sciences Section) under Grant GA-1707.

[3] On leave from The Economic Research Institute at the Stockholm School of Economics. Supported by a grant from The Bank of Sweden Tercentenary Fund.

[4] These discussions involved staff members from the Meteorology and Oceanography and the Psychology Departments of the University of Michigan and forecasters from the Detroit Metropolitan Airport Office of the National Weather Service.

[5] The framework for subjective probability prescribes that, in such a situation, a forecaster (whose utility function is linearly related to the scoring rule of concern) should act in such a way as to maximize his *expected* score. We assume that the scoring rule of concern is defined in such a way that a larger score is "better."

ever, a discussion of these questions, and their implications for subjective probability forecasting, is beyond the scope of this paper. We simply assume that situations *may* exist in which the introduction of second-order probabilities is appropriate from a meteorological point of view *and* reasonable from a probabilistic point of view.

In order to indicate the nature of the effect of additional information on second-order probabilities, consider an occasion on which, in the forecaster's judgment, the probability (of precipitation) p has a *uniform* distribution on the interval [0.2, 0.5]. That is, in the forecaster's judgment, the value of p lies in the interval between 0.2 and 0.5, inclusive, and all sub-intervals of this interval of equal length are equally likely to contain p. Then,

$$F(p)=\begin{cases} 1 , & p>0.5 \\ (10/3)\,(p-0.2), & 0.2\leqslant p\leqslant 0.5 \\ 0 , & p<0.2 \end{cases}$$

Note that the forecaster's *expected* judgment $E(p)$ is 0.35. Now, suppose that the forecaster receives an additional "item" of information, as a result of which he revises his judgment. In particular, suppose that his revised judgment is $F'(p)$, where

$$F'(p)=\begin{cases} 1 , & p>0.4 \\ 10(p-0.3), & 0.3\leqslant p\leqslant 0.4 \\ 0 , & p<0.3 \end{cases}$$

That is, in the forecaster's judgment, p is now *uniformly* distributed on the interval [0.3, 0.4]. Note that $E'(p) = 0.35$ [$=E(p)$]. Thus, this particular item of information has "sharpened up" the forecaster's judgment; however, the forecaster's expected judgment has not been changed. In general, of course, an item of information may "sharpen up" or "spread out" a forecaster's judgment and may or may not result in a change in his expected judgment.

3. The second-order model and strictly proper scoring rules

In this section, we investigate the effect of strictly proper scoring rules on the behavior of forecasters who subscribe to the second-order model for both a special and the general situation.

a. Special situation

Consider a two-state, i.e., "precipitation—no precipitation," situation in which the forecaster's judgment concerning the occurrence of precipitation is expressed in terms of a probability distribution $F(p)$ and in which the scoring rule of concern is the probability score PS (Brier, 1950), a strictly proper scoring rule (Murphy and Epstein, 1967). When his forecast probability (of precipitation) is r, the forecaster receives a score $PS(r)$, where

$$PS(r)=\begin{cases} 2(1-r)^2, & \text{if precipitation occurs} \\ 2r^2, & \text{if no precipitation occurs} \end{cases}$$

Then, the expected (probability) score is $E[PS(r,p)]$, or simply $E(PS)$, where

$$E(PS)=\int_0^1 PS(r,p)dF(p),$$

or

$$E(PS)=\int_0^1 \{p[2(1-r)^2]+(1-p)(2r^2)\}dF(p),$$

or

$$E(PS)=2(1-2r)\int_0^1 pdF(p)+2r^2\int_0^1 dF(p),$$

or

$$E(PS)=2(1-2r)E(p)+2r^2.$$

Now, differentiating $E(PS)$ with respect to r and setting the result equal to zero yields

$$-4E(p)+4r=0,$$

or

$$r=E(p).$$

Thus, the forecaster minimizes[6] his expected score by making his forecast r correspond to his expected judgment $E(p)$.

b. General situation

Consider a K-state situation in which the forecaster's judgment is expressed in terms of a K-dimensional probability distribution $F(\mathbf{p})=F(p_1,\cdots,p_K)$ defined on the $(K-1)$-dimensional simplex P, where

$$P=\{(p_1,\cdots,p_K)\,|\,p_k\geqslant 0, \sum_k p_k=1;\, k=1,\cdots,K\}.$$

Let $S_j(\mathbf{r})$ denote the score assigned by a strictly proper scoring rule S to a forecast $\mathbf{r}=(r_1,\cdots,r_K)$, where $r_k\geqslant 0$ and $\sum_k r_k=1$ $(k=1,\cdots,K)$, when state j obtains. Then, the forecaster's expected score is $E[S(\mathbf{r},\mathbf{p})]$, or simply $E(S)$, where

$$E(S)=\int_P S(\mathbf{r},\mathbf{p})dF(\mathbf{p}),$$

or

$$E(S)=\int_P [\sum_{j=1}^K p_jS_j(\mathbf{r})]dF(\mathbf{p}),$$

or

$$E(S)=\sum_{j=1}^K S_j(\mathbf{r})\int_P p_jdF(\mathbf{p}),$$

or

$$E(S)=\sum_{j=1}^K S_j(\mathbf{r})E(p_j). \tag{1}$$

[6] The probability score is defined in such a way that a smaller score is "better," and, as a result, the forecaster is concerned with *minimizing* his expected score. Note that $\partial^2[E(PS)]/\partial r^2>0$.

The form of $E(S)$, in (1), indicates that $E(p_j)$ plays the same role in the second-order model as p_j itself plays in the first-order model (Winkler and Murphy, 1968, p. 754; Staël von Holstein, 1970, p. 360; Murphy, 1970, pp. 919–920). Thus, the forecaster maximizes $E(S)$ by setting $\mathbf{r}$ equal to $E(\mathbf{p})$, i.e., by setting r_j equal to $E(p_j)$ for all j.[7]

4. Summary and conclusion

In this paper we have briefly considered the nature of second-order probabilities and described a second-order model of the subjective probability forecasting process. Then we have demonstrated that a forecaster's expected judgment[8] $E(\mathbf{p})$ plays the same role in a second-order situation in which his judgment is expressed in terms of a probability distribution $F(\mathbf{p})$ as the judgment $\mathbf{p}$ itself plays in the standard, first-order situation. In particular, strictly proper scoring rules, in the second-order situation, encourage the forecaster to make his forecast $\mathbf{r}$ correspond to his expected judgment $E(\mathbf{p})$.

In conclusion, we note that this result can be extended to situations in which higher-order models are appropriate (although such an extension is of theoretical rather than practical interest). In particular, for an Nth-order model, a forecaster whose judgment is expressed in terms of an Nth-order probability distribution maximizes his expected score, in the presence of a strictly proper scoring rule, by setting his forecast $\mathbf{r}$ equal to the $(N-1)$st order expected value of $\mathbf{p}$.

[7] If S is an *improper* scoring rule, then the optimal forecast, i.e., the forecast which maximizes $E(S)$, is the same as the optimal forecast under the first-order model except that p_j is (everywhere) replaced by $E(p_j)$.

[8] In the terminology of statistical decision theory, $E(\mathbf{p})$ is a *certainty equivalent* (Raiffa and Schlaifer, 1961, p. 178). That is, knowledge of the summary measure $E(\mathbf{p})$ is equivalent, from a decision-making point of view, to knowledge of the complete probability distribution $F(\mathbf{p})$.

REFERENCES

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

Murphy, A. H., 1970: The ranked probability score and the probability score: A comparison. *Mon. Wea. Rev.*, **98**, 917–924.

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

Raiffa, H., and R. Schlaifer, 1961: *Applied Statistical Decision Theory*. Boston, Harvard University, Graduate School of Business Administration, 356 pp.

Savage, L. J., 1954: *The Foundations of Statistics*. New York, Wiley, 294 pp.

Staël von Holstein, C.-A.S., 1970: A family of strictly proper scoring rules which are sensitive to distance. *J. Appl. Meteor.*, **9**, 360–364.

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