7.4.8 *Probability Forecasts for Multiple-Category Events*

Probability forecasts may be formulated for discrete events having more than two (yes vs. no) possible outcomes. These events may be nominal, for which there is not a natural ordering; or ordinal, where it is clear which of the outcomes are larger or smaller than others. The approaches to verification of probability forecasts for nominal and ordinal predictands differ, because the magnitude of the forecast error is not a meaningful quantity in the case of nominal events, but is potentially quite important for ordinal events. The usual approach to verifying forecasts for nominal predictands is to collapse them to a sequence of binary predictands. Having done this, Brier scores, reliability diagrams, and so on, can be used to evaluate each of the derived binary forecasting situations.

Verification of probability forecasts for multicategory ordinal predictands presents a more difficult problem. First, the dimensionality of the verification problem increases exponentially with the number of outcomes over which the forecast probability is distributed. For example, consider a $J = 3$-event situation for which the forecast probabilities are constrained to be one of the 11 values 0.0, 0.1, 0.2, . . . , 1.0. The dimensionality of the problem is not simply $33 - 1 = 32$, as might be expected by extension of the dimensionality for the dichotomous forecast problem, because the forecasts are now vector quantities. For example, the forecast vector (0.2, 0.3, 0.5) is a different and distinct forecast from the vector (0.3, 0.2, 0.5). Since the three forecast probabilities must sum to 1.0, only two of them can vary freely. In this situation there are $I = 66$ possible three-dimensional forecast vectors, yielding a dimensionality for the forecast problem of $(66 \times 3) - 1 = 197$. Similarly, the dimensionality for the four-category ordinal verification situation with the same restriction on the forecast probabilities would be $(286 \times 4) - 1 = 1143$. As a practical matter, because of their high dimensionality, probability forecasts for ordinal predictands primarily have been evaluated using scalar performance measures, even though such approaches will necessarily be incomplete.

For ordinal predictands, collapsing the verification problem to a series of $I \times 2$ tables will result in the loss of potentially important information related to the ordering of the outcomes. For example, the probability forecasts for precipitation shown in Figure 6.33 distribute probability among three MECE outcomes: dry, near-normal, and wet. If we were to verify the dry events in distinction to not dry events composed of both the near-normal and wet categories, information pertaining to the magnitudes of the forecast errors would be thrown away. That is, the same error magnitude would be assigned to the difference between dry and wet as to the difference between dry and near-normal.

Verification that is sensitive to distance usually is preferred for probability forecasts of ordinal predictands. That is, the verification should be capable of penalizing forecasts increasingly as more probability is assigned to event categories further removed from the actual outcome. In addition, we would like the verification measure to be strictly proper (see Section 7.4.7), so that forecasters are encouraged to report their true beliefs. The most commonly used such measure is the ranked probability score (RPS) (Epstein 1969b; Murphy 1971). Many strictly proper scalar scores that are sensitive to distance exist (Murphy and Daan 1985; Staël von Holstein and Murphy 1978), but of these the ranked probability score usually is preferred (Daan 1985).

The ranked probability score is essentially an extension of the Brier score (Equation 7.34) to the many-event situation. That is, it is a squared-error score with respect to the observation 1 if the forecast event occurs, and 0 if the event does not occur. However, in order for the score to be sensitive to distance, the squared errors are computed with respect to the *cumulative* probabilities in the forecast and observation vectors. This characteristic introduces some notational complications.

As before, let J be the number of event categories, and therefore also the number of probabilities included in each forecast. For example, the precipitation forecasts in Figure 6.33 have $J = 3$ events over which to distribute probability. If the forecast is 20% chance of dry, 40% chance of near-normal, and 40% chance of wet; then $y_1 = 0.2$, $y_2 = 0.4$, and $y_3 = 0.4$. Each of these components y_j pertains to one of the J events being forecast. That is, y_1, y_2, and y_3, are the three components of a forecast vector $\mathbf{y}$, and if all probabilities were to be rounded to tenths this forecast vector would be one of $I = 66$ possible forecasts $\mathbf{y}_i$.

Similarly, the observation vector has three components. One of these components, corresponding to the event that occurs, will equal 1, and the other $J - 1$ components will equal zero. In the case of Figure 6.33, if the observed precipitation outcome is in the wet category, then $o_1 = 0$, $o_2 = 0$, and $o_3 = 1$.

The cumulative forecasts and observations, denoted Y_m and O_m, are defined as functions of the components of the forecast vector and observation vector, respectively, according to

$$\mathrm{Y_m} = \sum_{j=1}^{m} \mathrm{y_j}, \quad \mathrm{m} = 1, \ldots, \mathrm{J}; \tag{7.46a}$$

and

$$\mathrm{O_m} = \sum_{j=1}^{m} \mathrm{o_j}, \quad \mathrm{m} = 1, \ldots, \mathrm{J}. \tag{7.46b}$$

In terms of the foregoing hypothetical example, $Y_1 = y_1 = 0.2$, $Y_2 = y_1 + y_2 = 0.6$, and $Y_3 = y_1 + y_2 + y_3 = 1.0$; and $O_1 = o_1 = 0$, $O_2 = o_1 + o_2 = 0$, and $O_3 = o_1 + o_2 + o_3 = 1$. Notice that since Y_m and O_m are both cumulative functions of probability components that must add to one, the final sums Y_J and O_J are always both equal to one by definition.

The ranked probability score is the sum of squared differences between the components of the cumulative forecast and observation vectors in Equation 7.46a and 7.46b, given by

$$\mathrm{RPS} = \sum_{m=1}^{J} (\mathrm{Y_m} - \mathrm{O_m})^2, \tag{7.47a}$$

or, in terms of the forecast and observed vector components y_j and $\mathrm{o_j}$,

$$\mathrm{RPS} = \sum_{m=1}^{J} \left[\left(\sum_{j=1}^{m} \mathrm{y_j} \right) - \left(\sum_{j=1}^{m} \mathrm{o_j} \right) \right]^2. \tag{7.47b}$$

A perfect forecast would assign all the probability to the single y_j corresponding to the event that subsequently occurs, so that the forecast and observation vectors would be the same. In this case, RPS $= 0$. Forecasts that are less than perfect receive scores that are positive numbers, so the RPS has a negative orientation. Notice also that the final ($m = J$) term in Equation 7.47 is always zero, because the accumulations in Equations 7.46 ensure that $Y_J = O_J = 1$. Therefore, the worst possible score is $J - 1$. For $J = 2$, the ranked probability score reduces to the Brier score, Equation 7.34. Note that since the last term, for $m = J$, is always zero, in practice it need not actually be computed.

EXAMPLE 7.7 Illustration of the Mechanics of the Ranked Probability Score

Table 7.6 demonstrates the mechanics of computing the RPS, and illustrates the property of sensitivity to distance, for two hypothetical probability forecasts for precipitation amounts. Here the continuum of precipitation has been divided into $J = 3$ categories, <0.01 in., $0.01 - 0.24$ in., and ≥ 0.25 in. Forecaster 1 has assigned the probabilities (0.2, 0.5, 0.3) to the three events, and Forecaster 2 has assigned the probabilities (0.2, 0.3, 0.5). The two forecasts are similar, except that Forecaster 2 has allocated more probability to the ≥ 0.25 in. category at the expense of the middle category. If no precipitation falls on this occasion the observation vector will be that indicated in the table. For most purposes, Forecaster 1 should receive a better score, because this forecaster has assigned more probability closer to the observed category than did Forecaster 2. The score for Forecaster 1 is $\text{RPS} = (0.2 - 1)^2 + (0.7 - 1)^2 = 0.73$, and for Forecaster 2 it is $\text{RPS} = (0.2 - 1)^2 + (0.5 - 1)^2 = 0.89$. The lower RPS for Forecaster 1 indicates a more accurate forecast.

If, on the other hand, some amount of precipitation larger than 0.25 in. had fallen, Forecaster 2's probabilities would have been closer, and would have received the better score. The score for Forecaster 1 would have been $\text{RPS} = (0.2 - 0)^2 + (0.7 - 0)^2 = 0.53$, and the score for Forecaster 2 would have been $\text{RPS} = (0.2 - 0)^2 + (0.5 - 0)^2 = 0.29$. Note that in both of these examples, only the first $J - 1 = 2$ terms in Equation 7.47 were needed to compute the RPS. ◊

Equation 7.47 yields the ranked probability score for a single forecast-event pair. Jointly evaluating a collection of n forecasts using the ranked probability score requires nothing more than averaging the RPS values for each forecast-event pair,

$$<\text{RPS}> = \frac{1}{n} \sum_{k=1}^{n} \text{RPS}_k. \tag{7.48}$$

TABLE 7.6 Comparison of two hypothetical probability forecasts for precipitation amount, divided into $J = 3$ categories. The three components of the observation vector indicate that the observed precipitation was in the smallest category.

	Forecaster 1		Forecaster 2		Observed	
Event	y_j	Y_m	y_j	Y_m	o_j	O_m
<0.01 in.	0.2	0.2	0.2	0.2	1	1
$0.01 - 0.24$ in.	0.5	0.7	0.3	0.5	0	1
≥ 0.25 in.	0.3	1.0	0.5	1.0	0	1

Similarly, the skill score for a collection of RPS values relative to the RPS computed from the climatological probabilities can be computed as

$$\mathrm{SS}_{\mathrm{RPS}} = \frac{<\mathrm{RPS}> - <\mathrm{RPS}_{\mathrm{Clim}}>}{0 - <\mathrm{RPS}_{\mathrm{Clim}}>} = 1 - \frac{<\mathrm{RPS}>}{<\mathrm{RPS}_{\mathrm{Clim}}>}. \tag{7.49}$$