AM 1.3

ADDITIONAL MEASURES OF SKILL FOR PROBABILISTIC FORECASTS

F. Wesley Wilson

National Center for Atmospheric Research

Research Applications Program

P.O. Box 3000, Boulder, CO 80307

Introduction

A measure of the skill of a forecast system is a statistic, which relates the observed performance of the system to its expected performance in similar circumstances. These discussions will focus on deterministic and probabilistic binary categorical forecasts, where the forecasted event can be verified either to occur (Obs=1) or to fail to occur (Obs=0). The deterministic forecast provides predictive value P=1 if the event is anticipated to occur and a predictive value P=0 otherwise. The probabilistic forecast provides a predictive probability P between 0 and 1, the probability that the event will occur.

The focus of these investigations is the comparison of the relationship between skill statistics for deterministic and probabilistic forecast systems. A probabilistic forecast system is confident if its forecasted probabilities are clustered near 0 and 1. A deterministic forecast system could be viewed as a very confident probabilistic forecast system. In this light, the skill statistics for probabilistic forecasts have meaning for deterministic forecasts. The literature pertaining to skill measures for deterministic forecasts provides many suggested skill statistics and considerable discussion relating to the value and frailties of these statistics. Can this knowledge provide insight into the value and frailties of measures of skill for probabilistic forecasts?

Skill is an elusive word, since it must be defined in terms of some philosophical predisposition of what constitutes a good and bad forecast. The viewpoints of the forecast provider and the forecast consumer may not agree in all regards. The viewpoint of the provider may involve institutional issues such as consistency, immunity from manipulation, and avoidance of extremes. The viewpoint of the consumer may be directed to such issues as economic value or correctness in special situations. The medial community has determined that the issues of false positives and false negatives deserve separate attention, resulting in a vector measure. The meteorological community has pressed for scalar measures. The trade-off is that the vector measure provides a more complete description of the skill, while the scalar is useful for setting decision thresholds and for selection of optimal strategies. The underlying fact is that the outcome space is partitioned by occurrence and non-occurrence, and that a forecast system may have different performance skill on these subsets of the outcome space. Mathematically, the binary performance of a deterministic forecast system is completely characterized by the (false positive, false negative) vector, and any condensation involves some loss of information regarding forecast performance.

Estimation of Skill Statistics

Skill statistics are estimated from verification trials, controlled situations where forecasts and outcomes are monitored. The results are used to estimate the value of the skill statistic. It is important to distinguish between a skill statistic and one of its estimators. The skill statistic has an intrinsic value, which describes the expected performance of the forecast system in similar situations. An estimator is a formula or procedure that is used to derive an estimated value of a statistic from the observed data. The efficiency of an estimator relates to rate at which the estimator converges to the value of the skill statistic, as the number of trials increases. A statistic may have both efficient and inefficient estimators. Most skill scores are described by formulas, which provide estimators for the associated skill statistic. In

some cases, the skill score differs from the efficient estimator for that skill statistic.

For deterministic forecasts of a binary event, the results of repeated trials are usually recorded in a 2 by 2 Scoring Contingency Table (Table 1). In each trial, the Forecast predicts either the occurrence of the event (P=1) or not (P=0), and the event is Observed either to verify (Obs=1) or not (Obs=0). The values A,B,C,D are the counts of the occurrences of the various observed states over the series of trials. Using # to denote the count of a set, we define A= #(P=1&Obs=1), B=#(P=1&Obs=0), C=#(P=0&Obs=1), and D=#(P=0&Obs=0). The margins of the Scoring Contingency Table are defined by M=A+B+C+D, M_1=A+C, and M_0=B+D. Normalization by these margins produces sample estimates for the probabilities indicated in Table 2. The intrinsic values of the conditional probabilities provide some skill statistics for the forecast system. In particular, **P**(P=1|Obs=0) is the probability of a false positive and **P(**P=0|Obs=1**)** is the probability of a false negative. From the viewpoint of the information in the Conditional Probability Contingency Table, forecast skill is improved if the forecast system is modified to decrease either of these conditional probabilities, without an increase in the other.

		Observation	
		1	**0**
Forecast	**1**	**A**	**B**
	0	**C**	**D**
		$\mathbf{M_1}$	$\mathbf{M_0}$ M

Table 1. The Standard Contingency Table for forecast verification.

Traditional Skill Scores are usually expressed as quotients of polynomials in the symbols from Table 1. There are four degrees of freedom since the skill scores are generated by the four parameters, {A, B, C, D}. Table 2 is obtained by normalization by M, and has three degrees of freedom. The information content of Table 2 is described without redundancy by the three parameters {**P**(P=1|Obs=1), **P**(P=1|Obs=0), **P**(Obs=1)}, cf. Marzban, 1998. All statistical properties of the performance of a forecast system, which are observable through these contingency tables, can be expressed algebraically in terms of the primitive parameters. Note that **P**(Obs=1) is a property of the weather (event frequency) and not a measure of the performance of the forecast system. There is latitude in choosing the descriptive parameters for these information spaces. Selection of the indicated parameters isolates the measures of forecast performance from the measure of event frequency. Skill statistics, which have dependence on the event frequency, are open to influence by the event frequency during the trial period. This can degrade the efficiency of statistical estimators.

Efficient estimation is an important reason to give preference to skill statistics, which can be expressed in terms of the parameters { **P**(P=1|Obs=1), **P**(P=1|Obs=0) }. These parameters are the basis for the classical Relative Operating Characteristic (ROC) diagram, e.g. Van Trees, 1968 and Mason, 1982. They are also featured in Doswell and Flueck, 1989.

An alternative approach to the estimation of skill statistics is to estimate these primitives separately, using efficient estimators for the primitives, and then to combine the estimates by the defining formula for the skill statistic. This differs from the traditional approach of using the companion skill score as the estimator. If all estimations are made using Table 1 from the same trial period, then the results of these approaches are identical. Different estimates are obtained when the primitives are estimated using trial periods of different lengths, perhaps chosen to reflect the efficiency of the various estimators. For example, *P(*Obs=1*)* could be estimated from a climatological database, and *P(*P=1|Obs=1*)* and *P(*P=1|Obs=0*)* estimated from a shorter trial period. Accurate estimations of the skill statistic might be obtained by this approach, but the estimator is different from the traditional skill score.

Some Standard Skill Statistics

There are many skill statistics for dichotomous, deterministic forecasts, e.g. the Hit Rate (HR) or proportion correct, the Conditional Bias, the Heidke Skill Statistic

Forecast		Observation	
		1	**0**
	1	**P(**P=1&Obs=1**)** = **P(**P=1|Obs=1**) P(**Obs=1**)**	**P(**P=1&Obs=0 **)** = **P(**P=1|Obs=0**) P(**Obs=0**)**
	0	**P(**P=0&Obs=1**)** = **P(**P=0|Obs=1**)P(**Obs=1**)**	**P(**P=0&Obs=0**)** = **P(**P=0|Obs=0**) P(**Obs=0**)**
		P(Obs=1**)**	**P(**Obs=0**)**

Table 2. The Conditional Probability Contingency Table.

(HSS) (Heidke, 1926), and the often-rediscovered Peirce Skill Statistic (PSS) (Peirce, 1884, Hanssen and Kuipers, 1965, and Flueck, 1987). Traditional skill scores are defined in terms of Table 1 and algebraic expressions that involve A, B, C, and D. The companion skill statistics are the limiting values of these skill scores as the number of trials becomes very large. Skill score formulas provide estimators for the skill statistics. This distinction is important. It provides a vocabulary for separate discussions of the intrinsic properties of the skill statistics and the accuracy of the estimations. In the following discussion, we use the notation ≈ to indicate "is estimated by".

HR is the skill measure that occurs to most practitioners initially. It is simply the proportion of time that the forecast is correct:

$$\begin{aligned} HR &= \mathbf{P}(P=1|Obs=1)\,\mathbf{P}(Obs=1) + \\ &\quad \mathbf{P}(P=0|Obs=0)\,\mathbf{P}(Obs=0) \\ &\approx (A + D) / M \end{aligned}$$

A forecast system, which is nearly perfectly correct, has HR=1; less obvious is the fact that a system with HR≈1 may have little skill. Peirce (1884) noted that for rare events (**P**(Obs=1)≈0), the system that always forecasts P=0 will also have HR≈1. HR is sensitive to the event frequency in the trial period.

PSS was originally introduced to measure the skill of the forecast system, when compared against a system of random forecasts that have the same marginal values. PSS is expressed in many forms. We shall use an algebraic form from Wilks, which conceals its genesis, but leads to the conditional probability expression from Doswell et al,1990:

$$\begin{aligned} PSS &= \mathbf{P}(P=1|Obs=1) - \mathbf{P}(P=1|Obs=0) \\ &\approx (AD - BC) / M_1 M_0 \\ &= (A/M_1) - (B/M_0) \end{aligned}$$

PSS has the benefits of having a simple interpretation in terms of forecast performance, of being efficiently estimated, and of being equitable, in the sense of Gandin and Murphy, 1992. This is strong evidence that PSS should be the preferred skill statistic, and vindicates Peirce's decision to designate it as "The numerical measure of success" (skill).

Forecast bias refers to a consistent offset of the forecasts from the correct forecast. For continuous predictands, bias is defined to be the expected or mean error between the forecasts and their verifications. That notion is not directly applicable to dichotomous forecasts, which provide no measure of the magnitude of error in an incorrect forecast. A consistent tendency to over- or under-forecast is another kind of bias. The most simplistic measure is the comparison of **P**(P=1) with **P**(Obs=1), the issue of whether or not the relative frequency of **P**=1 matches the relative frequencies of Obs=1. Measures of bias include

$$\Delta = \mathbf{P}(P=1) - \mathbf{P}(Obs=1)$$

$$\beta_1 = \mathbf{P}(P=1) / \mathbf{P}(Obs=1)$$

$$\beta_0 = \mathbf{P}(P=0) / \mathbf{P}(Obs=0)$$

$$\beta = (\beta_1 + \beta_0)/2$$

The best value for Δ is 0 and the best value for any of the ratios is 1; these are the values that define unbiased systems. Note that these are several different measures of the degree to which B differs from C. There is a simple strategy to improve the bias statistic, at the expense of degrading the skill of the forecast: Whenever B differs from C, forecasters can improve the β-statistics by keeping a running tally of their performance and applying the following rules

If B < C, forecast P=1 until B = C

If B > C, forecast P=0 until B = C

This practice, which can be added to any forecast system, provides an unbiased forecast system. We have introduced β to clarify the relationship of two of the most trusted skill statistics. We concur with the observation of Sanders (1963), that bias is a measure of system maturity, rather than system skill.

The HSS compares the HR of the trial period with the HR of a random forecast system. Like the PSS, there are many ways to express the algebra. We use the expression from Wilks, 1995:

$$HSS \approx \frac{2(AD - BC)}{(A+C)(C+D) + (A+B)(B+D)}$$

The numerator of HSS is twice the numerator of PSS. Additional algebra leads to the alternative expression:

$$HSS = PSS / \beta$$

Since β_T and β_F are positioned on opposite sides of 1, β often has values close to 1; so PSS $\approx$ HSS in many cases.

The list of standard skill statistics for categorical probabilistic forecasts is much shorter. There are two entries: the Brier Skill Score and the Reliability (Wilks, 1995). A forecast system is reliable if P(Obs=1|P=p) = p for every p, an appropriate generalization of the notion of an unbiased deterministic forecast. The Brier Score (Brier, 1950) is given by

$$BS = (1/n)\ \Sigma\ (P_i - Obs_i)^2$$

which is the mean-squared error (MSE) of the probabilities. Note that, if the forecast is very confident (deterministic), then BS = 1 – HR.

Skill Statistics for Probabilistic Forecasts

We shall now show that a generalization of the previous discussions provides extensions to probabilistic forecasts systems, of most of the skill statistics used for deterministic forecasts. The motivation for this is to provide more capabilities for the derivations and analyses of probabilistic forecast systems.

The MSE is the usual error residual, which is used for regression-based derivations. Thus the Brier Score holds an important position in the main stream of statistical forecasting. This is a mixed blessing. A major difficulty in the derivation of regression-based forecast models is the frequent occurrence of elongated depressions near the minimum, which makes it difficult to distinguish the optimal model. Indeed, there are many examples in deterministic forecasting where supplemental skill measures show substantial variation of forecast skill over the range of models, which have nearly equivalent MSE. One approach taken for deterministic models, is to select the model with highest PSS, among all those with similar MSE. An appropriate generalization of the PSS to probabilistic forecasts would extend this capability to the derivations of probabilistic forecasts.

The generalization to probabilistic forecasts is accomplished by the appropriate generalization of Table 2. In the deterministic case, instead of obtaining the entries A,B,C,D by counting, one could view these values as the results of summations:

$$A=\Sigma_1 P_i,\quad B=\Sigma_0 P_i,\quad C=\Sigma_1(1-P_i),\quad D=\Sigma_0(1-P_i)$$

where P_i is the i^{th} deterministic forecast, with value 0 or 1, and the summations are taken over the cases with Obs=1 and Obs=0, respectively. The extension to probabilistic forecasts is obtained by application of these formulas to the forecasted probabilities and defining the conditional probabilities

$$\mathbf{P(}P=1|Obs=1\mathbf{)} = A / M_1$$

$$\mathbf{P(}P=1|Obs=0\mathbf{)} = B / M_0$$

$$\mathbf{P(}P=0|Obs=1\mathbf{) =}\ 1\ \mathbf{-\ P(}P=1|Obs=1\mathbf{)}$$

$$\mathbf{P(}P=0|Obs=0\mathbf{)} = 1 - \mathbf{P(}P=1|Obs=0\mathbf{)}$$

Since the skill statistics are defined in terms of the conditional probabilities, their extensions to probabilistic forecasts are accomplished by applying the previous formulas. An important feature of this approach is that in the limit, as probabilistic forecasts become more confident, the values of all of these measures tend to the values of corresponding skill measures for deterministic forecasts.

An additional feature of this approach is that the probabilities of false positives and false negatives have interpretation in this context; so the skill of probabilistic forecasts has interpretation by these traditional measures.

Summary

The skill of categorical forecasts has been reviewed for deterministic, categorical forecasts. These measures have been interpreted from two viewpoints:

1. The meaning of the skill statistic in terms of conditional probabilities
2. The efficiency of the estimation of these statistics

These discussions provide additional information about the value of the Pierce Skill Statistic and provide a precise relationship between the Pierce and the Heidke statistics.

The Brier Statistic is identified for its role as the error residual in regression-based derivations of forecast equations. In the derivation of probabilistic forecast equations, there are similar limitations to those found in the derivations of deterministic equations. In the deterministic case, these limitations are frequently overcome by the introduction of supplemental skill statistics. Analogues of these supplemental statistics are introduced, through the generalizations of the appropriate conditional probabilities, enriching the capabilities for the derivations of probabilistic forecast equations. These additional skill statistics also provide a common ground for the comparison of the skill of deterministic and probabilistic forecast systems.

References

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. Mon Weather Rev., **78**, 1-3.

Brooks, H. B., and C. A. Doswell, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification, *Weather and Forecasting* **11**, 288-303.

Doswell, C.A. and J.A. Flueck, 1989: Forecasting and verifying in a field research project: DOPLIGHT '87. *Weather and Forecasting* **4**, 97-109.

Doswell, C.A., R. Davies-Jones, and D.L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting*, **5**, 576-585.

Doswell, C.A. and H.E. Brooks 1998: Budget cutting and the value of weather services. *Weather and Forecasting*, **13**, 206-212.

Flueck, J. A., 1987: A study of some measures of forecast verification. Preprints, *10th Conf. Probability and Statistics in Atmospheric Sciences*, Edmonton, Alberta. Amer. Meteor. Soc. 69-73.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Weather and Forecasting*, **13**, 753-763.

Murphy, A.H., 1973: A new vector partition of the probability score. *J. Appl. Met.*., **12**, 595-600.

Murphy, A.H. and H. Daan, 1985: Chapter 10: Forecast Evaluation, *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A.H. Murphy and R.W. Katz, Ed., Westview Press, Boulder and London.

Murphy, A.H. and R.W. Katz, Ed. 1985: *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, Westview Press, Boulder and London.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting, *Weather and Forecasting*, **8**, 281-293.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453-454.

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

Van Trees, H. L. 1968: *Detection, Estimation, and Modulation Theory*. Wiley, NY.

Wilks, D. 1995: Statistical Methods in the Atmospheric Sciences, Academic Press, NY