

The Expected Value of Frequency Calibration

ROBERT T. CLEMEN

University of Oregon

AND

ALLAN H. MURPHY

Oregon State University

It is possible to calibrate subjective probabilities using relative frequency information pertaining to a probability assessor's past performance. This procedure is known as frequency calibration and can be used to improve the quality of assessed probabilities. We develop a conceptual model of the probability assessment process and, on the basis of this model, show how to calculate the expected value of frequency calibration (EVFC) using standard Bayesian preposterior analysis. U.S. National Weather Service precipitation probability forecasts are used to illustrate the calculation of EVFC in the contexts of scoring rules and the familiar umbrella problem. © 1990 Academic Press, Inc.

When we ask for probability statements from experts, we like to think that those probability statements are "calibrated." Calibration can be thought of as long-run frequency calibration; over all those occasions when an individual said that the probability of an event was x , then on proportion x of those occasions the event actually occurred. We will say that an expert who is calibrated in this sense is "frequency-calibrated." This definition is consistent with the approach that has been taken by behavioral decision theorists (Lichtenstein, Fischhoff, & Phillips, 1982). Meteorologists have measured the performance of probability forecasters from this perspective, although they have used the term "reliability" rather than "calibration" (Murphy & Daan, 1985).

Another way to think about calibration is in a strict subjective sense. If

We thank Gary Carter of the NWS Techniques Development Laboratory for providing the data on which the empirical calculations of EVFC were based. Participants in the Decision Analysis Workshop at the Fuqua School of Business provided a number of helpful suggestions on an earlier version of this paper. Comments by Robert Winkler and two anonymous referees are gratefully acknowledged. This research was supported in part by the National Science Foundation under Grants IST-8600788 and ATM-871408. Inquiries or requests for reprints should be sent to R. T. Clemen, College of Business Administration, University of Oregon, Eugene, OR 97403.

a decision maker obtains a probability from an expert, the decision maker may wish to calibrate that probability subjectively. The adjustment that is made in this case would depend on the decision maker's careful assessment of the expert's ability. This assessment may be based in part on the expert's past performance (as in frequency calibration), but may also include other considerations such as adjustments for special circumstances associated with the event of interest and adjustments based on the decision maker's assessment of dependence between his or her own prior information and that of the expert. We call such adjustments "subjective calibration" following Morris (1977, 1983), Clemen (1986), and French (1986).

The topic of this paper is the expected value of frequency calibration (EVFC). That is, what is the value of conducting the frequency calibration procedure? Two immediate responses to this question come to mind, but neither adequately addresses all of the issues. The first is that EVFC must be zero for a Bayesian decision maker or forecaster, since such an individual would use all available information, including any historical data, to adjust a probability forecast (if necessary) before using it. This procedure, which may be either subjective or data-based, or a combination of the two, would satisfy the requirements of coherence; no further adjustment of the probability would be necessary. (By the same argument, the prior beliefs of a Bayesian decision maker must always be subjectively self-calibrated in order to preserve coherence.) It is nevertheless true that many real-world decision makers accept expert probability statements without questioning their calibration. EVFC thus represents the incremental expected value that the decision maker could gain through the frequency calibration process.

The second possible response is that EVFC is certainly positive. We know that most individuals who assess probabilities are poorly calibrated (Lichtenstein *et al.*, 1982). Even weather forecasters, although performing better in this regard than most individuals, generally exhibit some miscalibration (Ivarsson, Joelsson, Liljas, & Murphy, 1986; Murphy & Daan, 1984). Moreover, weather forecasters have the potential of improving their performance through a frequency calibration procedure (Clemen & Murphy, 1986b). Since EVFC is positive, the question should not be how much is frequency calibration worth, but why do we not routinely frequency calibrate all probability forecasts? The answer, of course, is that often the required data are unavailable.

EVFC can be interpreted in a variety of different ways depending on the specific circumstance. The fundamental interpretation, as previously mentioned, is that EVFC measures the incremental increase in expected value that a decision maker can anticipate from the frequency calibration process. If a decision maker takes an expert's statements at face value

when the assessments could be frequency calibrated, then the decision maker is ignoring valuable information. Another interpretation involves an expert whose response is one of a set of fixed values (e.g., 0.1, 0.2, . . . , 0.9), even though the expert's beliefs do not conform to this set. Here EVFC represents the potential gain that could be realized if the responses were not restricted in this way. Some probability assessments (notably those made by weather forecasters) are restricted to such a set of values. Informal observation of probability assessors suggests that many voluntarily restrict their assessments to "even" probabilities such as tenths, quarters, or thirds. The use of an assessment aid such as a probability wheel can help to eliminate this tendency (Spetzler & Staël von Holstein, 1975).

We will couch our analysis in terms of a weather forecaster who provides the probability of measurable precipitation during a well-defined future time period. This choice is made because our results are directly applicable to the evaluation of meteorological forecasts. Furthermore, meteorological data are abundant, providing an opportunity to calculate EVFC in a real-world situation; such examples are presented below. In spite of the weather forecasting context, our results are generalizable to any context in which (1) an individual (expert) provides probabilities and (2) data are available on the basis of which a decision maker can calibrate the expert's statements. For ease of exposition, we will maintain throughout the paper the distinction between the expert who generates the probabilities and the decision maker who uses them. EVFC is specific to the decision problem at hand and thus pertains to the decision maker rather than to the expert.

The work reported here is closely related to Epstein's (1966) treatment of uncertainty about a probability statement through beta density functions. Although the approaches are similar, Epstein viewed the problem in quality control terms, asking what kinds of data (about outcomes) would be consistent with certain beta density functions for an underlying proportion. In this paper, we are more concerned with an expected value question; given our current state of knowledge about a probability, how much would it be worth to us to improve the quality of the information?

In the next section we develop a probabilistic model of the forecasting process. We model the forecasting process as one in which the expert examines the current forecasting situation, considers it to be exchangeable with a set of similar situations previously observed, and thus provides the same probability that was assigned to the earlier situations. In the third section we show how EVFC can be calculated on the basis of the model, and we provide empirical examples based on meteorological data. The final section contains a discussion of limitations of the model and conclusions.

A MODEL OF THE PROBABILITY FORECASTING PROCESS

A Conceptual Approach

Suppose that an expert is formulating a probability of precipitation forecast for the next day. The subjective probability that precipitation will occur is formulated after considering an array of appropriate factors (climatological frequency of rain, "numerical" model output, current conditions, etc.). We postulate a process whereby the expert considers precipitation to be just as likely in the current situation as it has been on a number of previous occasions (e.g., see Sanders, 1963). It is helpful to visualize the process as classifying the situation at hand into a category for which the current situation is viewed as exchangeable with similar past situations. Figure 1 illustrates the concept. The boxes represent the categories, and the labels indicate in some way the chances of precipitation for days in the corresponding category. Non-numerical labels might be "rare," "uncommon," "toss-up," "uncertain," or "likely," to mention a few. (See Beyth-Marom, 1982, or Budescu & Wallsten, 1986, for a more complete list of possible terms.) Numerical labels also make sense. The expert could classify events into a category with the label "0.30" on the basis of a judgment that the long-run frequency of precipitation on these occasions is approximately 0.30.

We will let x_i denote the numerical label for category i . Let n_i denote the number of occasions contained in category i and r_i the number of occasions on which precipitation occurred. Now whenever the expert classifies a situation into category i , the frequency-calibrated probability of precipitation would be r_i/n_i . As more observations are gathered, r_i/n_i can be updated to reflect the recent information.

This conceptualization of the forecasting process is consistent with what we know about how people deal with uncertain situations. For example, Budescu and Wallsten (1986) report that people readily think in

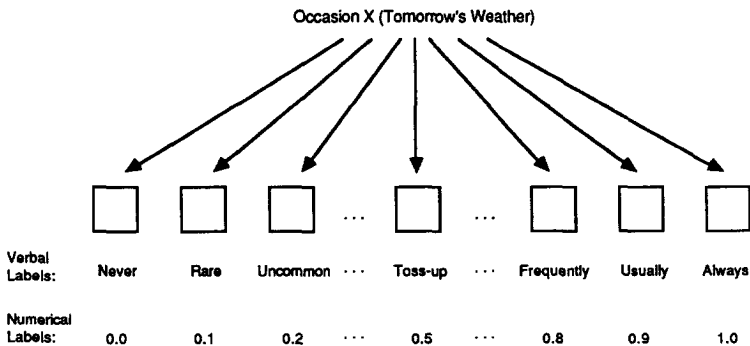


FIGURE 1

vague non-numerical terms. Zimmer (1983) and Murphy and Brown (1983) report that individuals typically distinguish among approximately five distinct probability categories. Furthermore, the process of frequency calibration as currently conceived (if not practiced) is consistent with our model.

This model is reminiscent of Kahneman and Tversky's (1972) representativeness heuristic, although the interpretation is not exactly what they had in mind. In their formulation, an individual assesses the probability of an object belonging to a category based on the extent to which the object fits the description of the stereotypical member of the category. In our case, the forecaster judges the extent to which an uncertain situation (as opposed to an object or event) resembles the "stereotypical uncertain situation" associated with the various categories and assigns the situation to the category which it most resembles. The frequency-calibrated probability of the event (precipitation) actually occurring depends on the overall frequency of occurrence for the specific category.

A Mathematical Model

This conceptual approach to the forecasting process leads naturally to a mathematical formulation. In order to complete the model, though, we require a few assumptions regarding the decision maker who uses the expert's assessments. We assume (1) that the decision maker accepts the process described above as a suitable representation of the way in which the expert generates probabilities; (2) that the decision maker views p_i , the probability of precipitation for events in category i , as an unknown parameter; and (3) that, after observing the expert's forecast x_i , the decision maker expresses his or her uncertainty about p_i through a probability distribution function $G_i(p_i)$. Coherence requires that the decision maker's subjective probability of precipitation in this case be Ep_i , the expected value of G_i . Finally, we will assume that the decision maker anticipates a reward for actions taken and subsequently realized outcomes. The decision maker's expected reward $V(\cdot, \cdot)$ is specific to the decision maker's particular decision problem and is a function of both p_i and the numerical value the decision maker uses to estimate p_i . For example, $V(x_i, p_i)$ is the expected reward when x_i is used as an estimate of p_i , whereas $V(r_i/n_i, p_i)$ is the expected reward when the decision maker uses the frequency-calibrated probability to estimate p_i .

Because p_i is unknown, the decision maker must consider the expected value of V taken over the distribution G_i . Given the above specification, the decision maker can calculate the expected reward for using the expert's probability statement x_i directly, incorporating into the expectation the uncertainty about p_i , as

$$E_{p_i}V(x_i) = \int_0^1 V(x_i, p_i) dG_f(p_i). \tag{1}$$

On the other hand,

$$E_{p_i}V(r_i/n_i) = \int_0^1 V(r_i/n_i, p_i) dG_f(p_i) \tag{2}$$

is the expected reward for using the frequency-calibrated probability. The p_i subscript on the expectation indicates that the integral is taken with respect to p_i ; for convenience of notation we will drop this subscript whenever the meaning is clear. Calculating the difference between the expected rewards in expressions (1) and (2) gives the expected value of frequency calibration for category i (EVFC):

$$\begin{aligned} \text{EVFC}_i &= \text{EV}(r_i/n_i) - \text{EV}(x_i) \\ &= \int_0^1 [V(r_i/n_i, p_i) - V(x_i, p_i)] dG_f(p_i). \end{aligned} \tag{3}$$

Subjective (rather than frequency) calibration would be equivalent to using E_{p_i} when the forecaster says x_i , and the expected value of subjective calibration (EVSC) would be

$$\begin{aligned} \text{EVSC}_i &= \text{EV}(E_{p_i}) - \text{EV}(x_i) \\ &= \int_0^1 [V(E_{p_i}, p_i) - V(x_i, p_i)] dG_f(p_i). \end{aligned} \tag{4}$$

Our formulation demonstrates clearly the difference between frequency and subjective calibration; subjective calibration implies the use of E_{p_i} , whereas frequency calibration implies the use of r_i/n_i . In many cases it will be reasonable to assume that $E_{p_i} = r_i/n_i$. For example, this assumption would be appropriate if G_i were a beta distribution with parameters r_i and n_i . Such a distribution is reasonable if the decision maker, having begun with an improper diffuse beta prior for p_i , has observed n_i days in category i , on r_i of which precipitation occurred.

EXAMPLES

In this section we consider some simple examples of reward functions to demonstrate the calculation of EVFC. We begin with strictly proper scoring rules, including the logarithmic and quadratic rules, and we calculate EVFC for U.S. National Weather Service (NWS) weather forecasters under the quadratic rule. We also consider EVFC for the umbrella problem (Katz & Murphy, 1987).

EVFC and Strictly Proper Scoring Rules

In this subsection we assume that the decision maker is rewarded via a

strictly proper scoring rule (Savage, 1971). EVFC in this case can be construed as the penalty that the decision maker pays for using probability assessments that are restricted to a fixed set of values. Denote the scoring rule by $S(x)$, where x represents the expert's stated probability. Specifically, let

$$S(x) = \begin{cases} S_1(x) & \text{if precipitation occurs,} \\ S_2(x) & \text{if precipitation does not occur.} \end{cases} \quad (5)$$

When p_i is the probability that precipitation occurs, the expected score for using the stated probability x_i is

$$V(x_i, p_i) = p_i S_1(x_i) + (1 - p_i) S_2(x_i). \quad (6)$$

Averaging over all possible values of p_i , the overall expected score for stating x_i is given by $EV(x_i)$:

$$\begin{aligned} EV(x_i) &= \int_0^1 [p S_1(x_i) + (1 - p) S_2(x_i)] dG_i(p) \\ &= Ep_i S_1(x_i) + (1 - Ep_i) S_2(x_i). \end{aligned} \quad (7)$$

Because S is a strictly proper scoring rule, $EV(x_i)$ is maximized by using Ep_i as the probability of precipitation when the expert reports x_i .

The decision maker's expected score for using the frequency-calibrated probability r_i/n_i is (7) with x_i replaced by r_i/n_i . Thus, the expected value of frequency calibration for category i is given by

$$\begin{aligned} EVFC_i &= EV(r_i/n_i) - EV(x_i) \\ &= Ep_i [S_1(r_i/n_i) - S_1(x_i)] + (1 - Ep_i) [S_2(r_i/n_i) - S_2(x_i)]. \end{aligned} \quad (8)$$

If $r_i/n_i = Ep_i$, this expression reduces to

$$EVFC_i = Ep_i [S_1(Ep_i) - S_1(x_i)] + (1 - Ep_i) [S_2(Ep_i) - S_2(x_i)]. \quad (9)$$

As an example, consider the logarithmic scoring rule:

$$S(x) = \begin{cases} \log(x) & \text{if precipitation occurs,} \\ \log(1 - x) & \text{if precipitation does not occur.} \end{cases} \quad (10)$$

The logarithmic scoring has been advocated by a number of Bayesian statisticians for theoretical reasons and because of its close tie with information theory (e.g., Good, 1952; Savage, 1971; and Bernardo, 1979, 1987). In this case, $EV(x)$ and $EVFC_i$ are given by

$$EV(x) = Ep_i \log(x) + (1 - Ep_i) \log(1 - x), \quad (11)$$

$$\begin{aligned} EVFC_i &= Ep_i \log(r_i/n_i x_i) \\ &\quad + (1 - Ep_i) \log[(n_i - r_i)/n_i (1 - x_i)]. \end{aligned} \quad (12)$$

Another example is the quadratic scoring rule, in which

$$S(x) = \begin{cases} -(1 - x)^2 & \text{if precipitation occurs,} \\ -x^2 & \text{if precipitation does not occur.} \end{cases} \quad (13)$$

The quadratic scoring rule was first introduced by Brier (1950) and is used by meteorologists to evaluate probabilistic weather forecasts (Murphy & Daan, 1985). Specifically, the quadratic scoring rule represents the squared forecast error when the probability is viewed as a forecast of the likelihood of occurrence of precipitation. The use of the quadratic rule by meteorologists and its squared error interpretation make it a particularly appealing scoring rule to consider in the context of EVFC.

EV(x) and EVFC_{*i*} for the quadratic scoring rule are given by equations (14) and (15):

$$EV(x) = -(Ep_i - x)^2 - Ep_i + (Ep_i)^2, \quad (14)$$

$$EVFC_i = (Ep_i - x_i)^2 - (Ep_i - r_i/n_i)^2. \quad (15)$$

Moreover, if $Ep_i = r_i/n_i$, then (15) becomes simply

$$EVFC_i = (r_i/n_i - x_i)^2. \quad (16)$$

Finally, we have considered only the *i*th category. Aggregating over all *k* categories, the overall EVFC is simply a weighted average of the EVFCs for each individual category:

$$EVFC = N^{-1} \sum_{i=1}^k n_i EVFC_i, \quad (17)$$

where $N = \sum n_i$. Equation (17) is equivalent to a fully Bayesian approach in which the next forecast situation belongs to category *i* with unknown probability q_i and $Eq_i = n_i/N$.

Our model can be used to analyze probability of precipitation forecasts to estimate the improvement in expected or average score resulting from frequency calibration. Because a weather forecaster's performance is evaluated in part on the basis of such scores, we have a situation in which the forecaster faces the reward function *V*. Thus, in this context, the expert and the decision maker are one and the same, and EVFC can be interpreted as the penalty that the weather forecaster pays for having to restrict the probability forecast to a specific set of values.

Figure 2 shows four calibration curves based on weather forecasters' assessments of the probability of precipitation (PoP). These data are the same as those employed by Clemen and Murphy (1986a,b), and cover the period from April 1972 through September 1983. NWS forecasters gen-

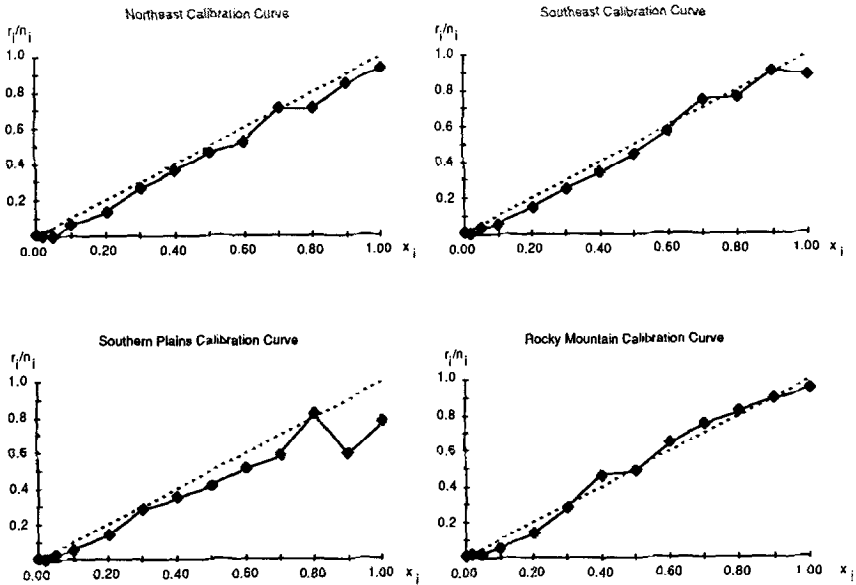


FIGURE 2

erate PoP forecasts twice each day for several different lead times. Furthermore, meteorologists traditionally analyze warm (April through September) and cool (October through March) seasons separately. In this example we consider only forecasts made during the warm season in conjunction with the so-called 0000 UTC cycle for a period of 12 to 24 h into the future (which corresponds to a 12-h daytime period).

For each of the four geographical areas, we aggregated forecasts over four NWS offices (Table 1). Because forecasts from individual NWS offices are generally made by many different forecasters, aggregation of the

TABLE 1
NWS OFFICES AND AREAS DEFINED IN TERMS OF THE OFFICES FOR WHICH EVFC WAS CALCULATED IN THIS STUDY

Area	Offices
Northeast	Albany, NY
	New York, NY
Southeast	Asheville, NC
	Birmingham, AL
Southern Plains	Amarillo, TX
	Oklahoma City, OK
Rocky Mountain	Boise, ID
	Great Falls, MT
	Boston, MA
	Philadelphia, PA
	Atlanta, GA
	Columbia, SC
	Dallas/Ft. Worth, TX
	Wichita, KS
	Denver, CO
	Salt Lake City, UT

data over areas with relatively homogeneous meteorological and climatological regimes would not appear to compromise the results of this study in any material way.

Table 2 provides the data that are necessary for the determination of EVFC, and we calculate EVFC for the quadratic scoring rule. The quadratic scoring rule was used because of its widespread application in meteorology. Furthermore, its use here facilitates comparisons with Clemen and Murphy's (1986b) empirically determined improvement in

TABLE 2
DATA, EVFC, AND EMPIRICAL IMPROVEMENT FOR THE FOUR NWS AREAS

Forecast probability	Area			
	Northeast	Southeast	Southern Plains	Rocky Mountain
0.00	0.01 1752	0.01 1521	0.01 2066	0.01 1618
0.02	— 0	0.00 13	0.00 30	0.02 112
0.05	— 0	0.03 94	0.03 266	0.02 697
0.10	0.06 1078	0.05 990	0.06 1759	0.06 1397
0.20	0.13 811	0.15 937	0.13 1260	0.13 1128
0.30	0.26 587	0.25 923	0.38 634	0.28 795
0.40	0.36 371	0.34 564	0.35 326	0.46 370
0.50	0.46 332	0.45 482	0.41 190	0.48 289
0.60	0.52 253	0.57 386	0.51 130	0.64 250
0.70	0.70 175	0.74 203	0.58 88	0.75 118
0.80	0.70 185	0.76 135	0.81 70	0.82 90
0.90	0.84 130	0.90 73	0.59 29	0.89 46
1.00	0.93 202	0.88 90	0.77 22	0.94 53
EVFC	0.0022	0.0022	0.0031	0.0019
Empirical improvement	0.0039	0.0014	0.0023	0.0014
<i>n</i>	784	1089	1466	1486
<i>t</i>	3.04	1.30	2.14	1.96

Note. In each cell the upper number is r_i/n_i and the lower number is n_i . For purposes of calculating EVFC, we set $Ep_i = r_i/n_i$. The last three lines of the table are based on Clemen and Murphy (1986b).

expected quadratic score due to frequency calibration (also included in Table 2). Recall that EVFC represents the anticipated improvement in expected score due to frequency calibration. For the Northeast area, the actual improvement found by Clemen and Murphy was slightly greater than EVFC, but for the Southeast, Southern Plains, and Rocky Mountain areas the actual improvement was slightly less than EVFC.

The t -statistics reported at the bottom of Table 2 can be used to test the null hypothesis that the empirical improvement due to frequency calibration is zero or negative versus the alternative hypothesis that the improvement is positive. The only insignificant value is for the Southeast area. Thus, we have some empirical statistical support for the positive value of the frequency calibration procedure. Furthermore, although the reported improvements do not appear large, they amount to approximately twice the average annual improvement in forecasting accuracy over the past 18 years (Murphy & Sabin, 1986).

EVFC and the Umbrella Problem

In this subsection we consider a problem in which a decision maker uses an expert's probability as an input in solving a decision problem. Thus, in contrast to the previous subsection, here EVFC is construed as the decision maker's penalty for not calibrating the expert's probability assessments when they should be calibrated.

The decision problem we address is the familiar umbrella problem, for which a decision tree is presented in Fig. 3. In this problem the decision maker's alternatives are to take an umbrella on an outing or to leave it at home. Taking the umbrella incurs cost C (regardless of the weather conditions). If the umbrella is left, the outcome depends on the weather; zero expense is associated with not taking the umbrella on a day without rain, but loss $L > C$ is sustained on a rainy day with no umbrella. The expected-expense-minimizing solution is to take the umbrella when $P(\text{rain}) > C/L$.

Frequency calibration would have value in this setting only if the fre-

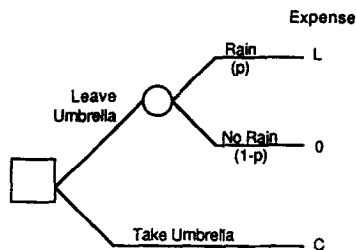


FIGURE 3

quency-calibrated probability led to a different decision than the uncalibrated probability. For example, suppose that the stated probability of rain is $P(\text{rain}) = x_i$, and

$$r_i/n_i < C/L < x_i. \tag{17}$$

Assume for convenience that $r_i/n_i = Ep_i$. If x_i were used as the probability of rain, the decision would be to take the umbrella, whereas using $P(\text{rain}) = r_i/n_i$ would mean not taking the umbrella. In this case

$$\text{EVFC}_i = C - Lr_i/n_i. \tag{18}$$

Likewise, if

$$x_i < C/L < r_i/n_i, \tag{19}$$

then

$$\text{EVFC}_i = Lr_i/n_i - C. \tag{20}$$

Several such forecast values may exist for which the decision under frequency calibration differs from the decision under the stated probability. The overall expected value of frequency calibration is given by:

$$\text{EVFC} = N^{-1} \sum_{i=1}^k n_i [(r_i/n_i)L - C] \gamma(x_i), \tag{21}$$

where

$$\gamma(x_i) = \begin{cases} 1 & \text{if } x_i < C/L < r_i/n_i, \\ -1 & \text{if } r_i/n_i < C/L < x_i, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

We can use our PoP forecast data to demonstrate the calculation of EVFC in a problem of this nature. Consider, for example, a situation in which a decision must be made regarding the protection of freshly-poured concrete. Once the concrete is poured, the options are to implement a one-time protective measure at cost C or to risk damage due to precipitation, incurring loss L with probability p . In such a situation, C/L is believed to lie in the range from 0.20 to 0.40 (Liljas, 1984). For each of the four areas, EVFC is graphed in Fig. 4 as a percentage of L for values of C/L between 0.20 and 0.40. In all four areas, $r_{0.3}/n_{0.3} < 0.3$, and thus for each area EVFC is positive for values of C/L between $r_{0.3}/n_{0.3}$ and 0.3. However, $r_{0.4}/n_{0.4} < 0.4$ only for the Northeast, Southeast, and Southern Plains areas. For each of these three areas EVFC is positive for values of C/L between $r_{0.4}/n_{0.4}$ and 0.4. (For the Rocky Mountain area EVFC is positive for C/L between 0.40 and 0.46.) When C/L is close to 0.3 and L is moderate (on the order of \$100,000), EVFC can be as high as \$700 for

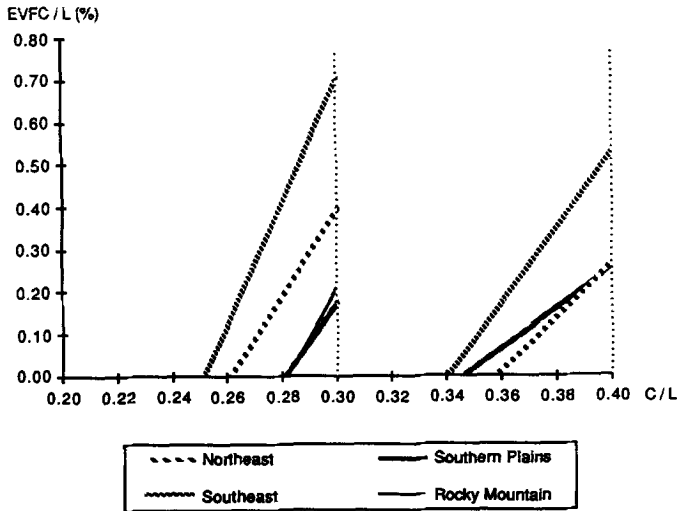


FIGURE 4

a single decision. Considering the multitude of such decisions that may rely on weather forecasts, it would appear that calibrating weather forecasts could lead to substantial benefits.

SUMMARY AND CONCLUSIONS

We have presented an approach to estimating a priori the expected value of frequency calibrating probabilities from an expert. The conceptual model is consistent with the way we think about frequency calibration, and the mathematical elaboration of the model is based on Bayesian preposterior analysis (Raiffa & Schlaifer, 1961). Examples involving weather forecasts have shown how to calculate EVFC in situations involving different types of payoff functions. EVFC for the meteorological experts is fairly low because these experts are, as indicated, better calibrated than many probability assessors. For less well-calibrated experts whose assessments are used in comparable decision situations, EVFC would be greater.

Our approach, however, does possess some limitations. First, the model tacitly assumes that the expert's assessment ability is stationary. That is, the expert cannot "second-guess" his or her tendency to over- or underestimate probability p_i in a conscious self-calibration effort. Furthermore, it is not clear that experts could perform frequency calibration for their own assessments as an explicit separate step after making the original assessments. The anticipation of doing so could affect the assessments in the first place.

Another limitation is that the model relies on a sequence of exchangeable events. This assumption implies that the expert cannot learn about the real-world system over time and become a better forecaster (i.e., by learning to discriminate more accurately between wet and dry periods). Such an assumption is obviously unreasonable for an expert who is becoming newly acquainted with a particular field. However, for established probability forecasters, such learning may be quite slow, and so this assumption may be realistic. Clemen and Murphy's (1986b) analysis of a large sample of PoP forecasts demonstrated that stable biases do exist over time. If some learning is occurring over time, it may be possible to model the learning process (DeGroot, 1980). An ad hoc solution would be to give more weight to recent observations rather than using the equally weighted r_i/n_i .

Even though much is known about the extent to which probability assessments are calibrated, the process of frequency calibration appears to be uncommon. The reason seems to be that in many cases adequate data sets do not exist. Our analysis as described requires some data simply to calculate EVFC. As a first step, though, a decision maker could calculate EVSC on the basis of subjectively assessed distributions G_i . If EVSC is high enough, then it might be reasonable to begin the data collection process necessary for frequency calibration. Of course, the G_i distributions can be updated using new data, and sequential EVFC calculations can be performed, providing an ongoing check on the value of the frequency-calibration program.

We have argued that EVFC can be thought of as a potential increase in expected value that is foregone when a decision maker uses probability assessments directly when the assessments should be adjusted. Our perspective is perhaps best seen in the spirit of Simon's (1957) bounded-rationality approach to human decision making. In general, the Bayesian approach described here can be used to measure the value of additional information processing and thus may be quite useful in studying situations where such processing is costly. In our case, some costs are involved in calibrating forecasts; calculation of EVFC, although costly itself, can show whether a routine program of frequency calibration would be worth the effort and resources required.

REFERENCES

- Beyth-Marom, R. (1982). How probable is probable? Numerical translation of verbal probability expressions. *Journal of Forecasting*, 1, 257-269.
- Bernardo, J. (1979). Expected information as expected utility. *Annals of Statistics*, 7, 686-690.
- Bernardo, J. (1987). Approximations in statistics from a decision-theoretical viewpoint. In R. Viertl (Ed.), *Probability and Bayesian statistics* (pp. 53-60). New York: Plenum.

- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1-3.
- Budescu, D. V., & Wallsten, T. S. (1986). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, 36, 391-405.
- Clemen, R. T. (1986). Calibration and the aggregation of probabilities. *Management Science*, 32, 312-314.
- Clemen, R. T., & Murphy, A. H. (1986a). Objective and subjective precipitation probability forecasts: Statistical analysis of some interrelationships. *Weather and Forecasting*, 1, 56-65.
- Clemen, R. T., & Murphy, A. H. (1986b). Objective and subjective precipitation probability forecasts: Improvements via calibration and combination. *Weather and Forecasting*, 1, 213-218.
- DeGroot, M. H. (1980). Improving predictive distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 385-395). Valencia: University Press.
- Epstein, E. S. (1966). Quality control for probability forecasts. *Monthly Weather Review*, 94, 487-494.
- French, S. (1986). Calibration and the expert problem. *Management Science*, 32, 315-321.
- Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, 14, 107-114.
- Ivarsson, K.-I., Joelsson, R., Liljas, E., & Murphy, A. H. (1986). Probability forecasting in Sweden: Some results of experimental and operational programs at the Swedish Meteorological and Hydrological Institute. *Weather and Forecasting*, 1, 136-154.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430-454.
- Katz, R. W., & Murphy, A. H. (1987). Quality/value relationship for imperfect information in the umbrella problem. *The American Statistician*, 41, 187-189.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). New York: Cambridge University Press.
- Liljas, E. (1984). Benefits resulting from tailored very-short-range forecasts in Sweden. *Nowcasting II; Mesoscale observations and very-short-range weather forecasting* (ESA Report No. sp-208, pp. 503-507). Noordwijk, The Netherlands: European Space Agency.
- Morris, P. A. (1977). Combining expert judgments: A Bayesian approach. *Management Science*, 23, 679-693.
- Morris, P. A. (1983). An axiomatic approach to expert resolution. *Management Science*, 29, 24-32.
- Murphy, A. H., & Brown, B. G. (1983). Forecast terminology: Composition and interpretation of public weather forecasts. *Bulletin of the American Meteorological Society*, 64, 13-22.
- Murphy, A. H., & Daan, H. (1984). Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment. *Monthly Weather Review*, 112, 413-423.
- Murphy, A. H., & Daan, H. (1985). Forecast evaluation. In A. H. Murphy & R. W. Katz (Eds.), *Probability, statistics, and decision making in the atmospheric sciences* (pp. 379-437). Boulder, CO: Westview Press.
- Murphy, A. H., & Sabin, T. E. (1986). Trends in the quality of National Weather Service forecasts. *Weather and Forecasting*, 1, 42-55.

- Raiffa, H., & Schlaifer, R. O. (1961). *Applied statistical decision theory*. Boston: Harvard University Press.
- Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, 2, 191-201.
- Savage, L. J. (1971). The elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66, 783-801.
- Simon, H. A. (1957). *Models of man*. New York: Wiley.
- Spetzler, C. S., & Staël von Holstein, C.-A.S. (1975). Probability encoding in decision analysis. *Management Science*, 22, 340-358.
- Zimmer, A. C. (1983). Verbal versus numerical processing of subjective probabilities. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 159-182). Amsterdam: North-Holland.

RECEIVED: April 20, 1988.