ARPA ORDER NO. 189-1

RM-6299-ARPA
JUNE 1970

PROBABILISTIC FORECASTS AND REPRODUCING SCORING SYSTEMS

Thomas A. Brown

prepared for
ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

ARPA ORDER NO. 189-1

RM-6299-ARPA
JUNE 1970

PROBABILISTIC FORECASTS AND REPRODUCING SCORING SYSTEMS

Thomas A. Brown

This research is supported by the Advanced Research Projects Agency under Contract No. DAHC15 67 C 0141. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of Rand or of ARPA.

Rand
SANTA MONICA, CA. 90406

Published by The RAND Corporation

PREFACE

This memorandum was produced as part of a program in group judgment technology conducted for the Behavioral Sciences Office of the Advanced Research Projects Agency. For many military problems, the best information available is the judgment of knowledgeable individuals. This is especially true in the assessment of long-range technological developments, and the evaluation of long-range future threats. Thus the military has an important stake in ensuring that the procedures used for obtaining judgments are adequately designed to elicit the most accurate estimates possible from the community of experts.

By their very nature, these estimates are uncertain; therefore, it seems reasonable that they should be couched in probabilistic terms. This memorandum discusses ways in which the incentive system imposed on experts may be structured so as to induce the best possible performance in probabilistic forecasts, and briefly touches on ways in which such forecasts may be combined into consensus forecasts.

Related material may be found in RM-5888-PR, RM-5957-PR, RM-6115-PR, and RM-6118-PR.

SUMMARY

It seems worthwhile, when asking groups of experts for forecasts of political, economic, or military events, to occasionally require them to put their forecasts in probabilistic terms.

The advantages of such an approach are as follows:

(a) It provides a concise expression of subjective uncertainty.

(b) It provides an operational self-rating as to the degree of confidence to be placed in the forecast.

(c) It is readily usable in decision-theoretic models.

(d) It is easily combined with other forecasts couched in similar terms.

To induce accurate forecasts, it seems reasonable to reward these experts by a scheme related to the extent to which their forecasts "come true." Such a scheme must be carefully chosen lest it contain built-in incentives to exaggerate or to understate probabilities. Systems which are free of distorting incentives are called "reproducing scoring systems"; such systems exist for both predictions relating to discrete alternatives and predictions couched in terms of continuous distributions. Reproducing scoring systems have been applied to weather forecasts and classroom tests. Their application to political, military, and economic forecasts, however, may give rise to the following problems:

(a) Forecasters may not attempt to maximize their expected gains.

(b) Conflict of interest can pollute such a system as it can any other.

(c) A great many forecasts must be considered before a reproducing scoring system will reliably distinguish accurate from inaccurate forecasters.

Examination of the advantages and disadvantages of various probabilistic scoring systems is a fit subject for experimental investigation; such experiments are now being designed and will be reported in due course.

CONTENTS

1. WHY PUT FORECASTS IN PROBABILISTIC TERMS?

Almost every decision we make, in private or in public life, is implicitly based on forecasts of future events and conditions. Naturally we are willing, therefore, to put considerable effort into various types of forecasting and to consult experts in various fields in order to improve our knowledge of the likely course of events in the future. Everyone recognizes that the future is uncertain, however, and thus it would seem very natural for forecasts to be generally cast in probabilistic terms. For example, in forecasting the results of the 1968 presidential election one might have said "Nixon .60; Humphrey .38; Wallace .02," or some similar apportionment of probabilities. This quantitative type of forecast has several obvious advantages:

(a) It provides a concise expression of subjective uncertainty.

(b) It provides an operational self-rating as to the degree of confidence to be placed in the forecast. Someone who has not studied the electorate carefully will be more likely to smear his probability assignment over all possible alternatives than someone who has a deeper knowledge of the forces at work.

(c) It provides a forecast which is readily usable by those using the tools of decision theory in choosing between alternative courses of action.

(d) It provides a forecast which may be easily combined with other people's forecasts couched in the same terms. There are many ways in which this may be done. The most obvious method is simply to average the probabilities ascribed to each alternative by the set of forecasters. Edwards and Phillips[8] have found that in the cases they examined subjects tended to give more accurate probability estimates if asked for odds rather than probabilities. In a personal communication, Edwards has suggested using the mean of the logarithms of the odds as a good way of combining a collection of probability estimates. A method of Eisenberg and Gale[9,16] takes as the consensus the odds which would be arrived at if the forecasters were allowed to place bets on the alternatives as at a race track. These three methods may give quite different results: If two forecasters ascribe probabilities to two alternatives, and one ascribes probabilities (.5, .5) while the other ascribes probabilities (.1, .9), then the various consensus algorithms give the following results:

	Alternative 1	Alternative 2
Parimutuel method	.5	.5
Average probability method	.3	.7
Mean log odds method	.25	.75

The question of which consensus technique is most appropriate is analogous to the question of which measure of

central tendency (mean, mode, or median) for a body of data is most appropriate. But it is a question which can be addressed, experimentally and theoretically, with precision. Contrast this with the problem of creating a consensus out of a handful of the essays normally produced as forecasts by political pundits.

These advantages seem so overwhelming that it is somewhat surprising how few forecasts are actually put in terms of explicit probabilities. There must, therefore, be factors which militate against putting forecasts in such terms. Some of these factors may be the following:

(a) Forecasters may be highly knowledgeable in their field of expertise, but ignorant of the language of probabilities. The world is full of otherwise intelligent people who think that chuck-a-luck is a fair game, and who have very little skill at expressing their expectations in terms of probabilities. Numerous psychological experiments have exposed the difficulty with which most people handle probability concepts. For example, Edwards[7] has found that most people do not "understand" Bayes' theorem, in the sense that they will not change prior subjective probabilities as much as they should in response to new evidence; that is, they are too conservative. On the other hand, some unpublished experiments of Dalkey and Cochran indicate that students overstate their confidence that given pairs of words are synonyms or antonyms.

(b) Desire for vagueness may afflict some forecasters who, like clever fortune-tellers, wish to make "prophecies" which appear to have content but which actually will be "proved correct" independently of the actual course of events. Vagueness may also be a bureaucratic necessity at times. For example, a National Intelligence Estimate ordinarily represents a consensus between State, the CIA, DIA, AEC, and FBI. Ambiguous verbal formulas are sometimes required in order to produce a document on which all these agencies can agree.

(c) Users may pressure forecasters to come out flatly behind one alternative or another. To some extent this may be a maneuver designed to counter the forecaster's desire for vagueness; on the other hand, it may be related to a failure to realize that an honest probabilistic forecast gives considerably more information about a forecaster's views than simply naming the single alternative which he considers most likely. Furthermore, good planning should take into account all reasonable contingencies rather than just the single most likely one.

(d) Epistemological difficulties arise when you speak of the "probability" of an inherently unique event. The frequentist notion of probability propounded by logical positivists such as Von Mises can give meaning to concepts such as "The probability that Nixon will win the 1968 presidential election" only by means of a very tortured con-

struction. The Von Mises notion of probability is, in fact, not as mathematically precise as it might at first appear[14] but a version of it can be made precise with the aid of recursive function theory (see, for example, DiPaola.[6]) It would be false sophistication, however, for one to be inhibited from using the language of probability in the way in which it is used daily by actuaries and gamblers simply because a particular approach to the foundations of probability has difficulty in accounting for such usage.

(e) Desire for credit, in a sense the opposite to the desire for vagueness, may influence forecasters to make specific rather than probabilistic predictions. If you predict Nixon will win the election, you will be proved right or wrong by the event. If you say Nixon has a .6 chance of winning, what will the actual event prove about your forecast?

The first three factors above are, of course, problems with forecasters and users rather than problems with forecasts. These difficulties can be removed by merely seeing to it that human beings become more intelligent and more virtuous. The fourth difficulty, the epistemological question, is a deep, complex matter about which reasonable men may differ. It is a problem of rather abstract philosophy and to deal with it adequately would require a lengthy tract. The fifth problem is the subject of this paper; how should you apportion credit to probability forecasters

after the fact? We shall champion a class of scoring systems called "reproducing scoring systems."

2. WHY REPRODUCING SCORING SYSTEMS?

2.1 What Are Reproducing Scoring Systems?

Let us suppose that an expert reports to you probabilities $r_1, r_2, \ldots, r_n$ that each of n mutually exclusive alternatives is going to take place (and at least one of them must take place, so that if the expert is internally consistent $\Sigma r_i = 1$). You do not pay him at once for this forecast, but wait until one of the alternatives (say, alternative i) actually takes place, and then pay him an amount

$$f_i(r_1, r_2, \ldots, r_n)$$

The functions f_i ought to be chosen in such a way that, if the expert truly believes that $p_1, p_2, \ldots p_n$ are the probabilities of the event in question, then his perception of the expected payoff from his prediction

$$\sum_i p_i f_i(r_1, r_2, \ldots, r_n)$$

will be a maximum for $r_1 = p_1$, $r_2 = p_2$, etc. If a reward system has this property, it is called a "reproducing scoring system" because it gives the expert an incentive to report his true beliefs (to "reproduce" them).

We insist, therefore, that the maximum at $r_i = p_i$ be a strict maximum (e.g., we do not call $f \equiv 0$ a "reproducing

scoring system").

2.2 Why Use Reproducing Scoring Systems?

There are two basic reasons to use reproducing scoring systems:

(a) So that the forecaster will provide you with a true report of his judgments. Even if there is no overt, explicit score being kept, the forecaster will perceive that somehow or other his forecast will affect his personal future in some way (otherwise he will be unwilling to put any effort into arriving at a reasonable forecast); the link which he perceives between his forecast and his personal fortunes is bound to have a strong influence on both the amount of effort he puts into his forecast and also on the content of the forecast. The use of a reproducing scoring system is an attempt to explicitly structure this link in such a way that it will influence the forecaster to make a sincere effort to determine what is likely to happen, and then accurately report his view to the user.

(b) So that, over the long term, you can sort out better forecasters from poorer ones. It appears that today forecasters are judged more on the quality of their writing and on their political sense (what do people want to hear?) than on the quality of their forecasts. A systematic scoring system would provide an arena in which good forecasters could excel on the basis of how accurately they

perceived the future rather than on the basis of "styling" and "catching the word market." As we shall see below, however, there is a subtle conflict between this use of reproducing scoring systems and the use given above.

2.3 Some Horrible Examples

The reader at this point may feel that the arguments in favor of eliciting probabilistic forecasts and scoring them in some explicit way are valid, but that any "reasonable" scoring system "will do," and that it is not necessary to use a reproducing scoring system. In this section we will present three examples of "reasonable" scoring systems which are not reproducing scoring systems, and point out the distortions in forecasts which which they will encourage.

(a) The simplest possible scoring system is to simply take

$$f_i(r_1, r_2, \ldots, r_n) = r_i$$

This scoring system means that the forecaster will maximize his expected score by picking the outcome he considers most likely, and claiming that he is certain it will take place. That is,

$$r_i = 1 \qquad \text{if } p_i = \max_j \{p_j\}$$

$$= 0 \qquad \text{otherwise.}$$

Over a period of time, then, this scoring system will tend to benefit the incautious forecaster and penalize the forecaster who reports his true feelings.

(b) In some of Norm Dalkey's experiments with the Delphi method, the respondents produced subjective probability distributions for a certain quantity whose true value was known to the experimenters. The question arose how to rate these probability distributions for accuracy, and one suggestion was the following: given that the true value was x and the probability density function submitted by a respondent was r(y)dy, pick a suitable function ρ and let the respondent's "score" be

$$f(r(.)) = 1 - \int_D \rho(|x-y|)r(y)dy$$

where D is the domain over which r(y) is defined. Let us suppose that p(x)dx is the probability density function which the respondent "really" has in mind. Then he will perceive his expected score as being

$$E(f(r(.))) = 1 - \int_D r(y) \{\int_D \rho(|x-y|)p(x)dx\}dy$$

Note that the quantity inside the curly brackets is a function of y alone. Thus the respondent may maximize his expected score by concentrating all the mass of r(y)dy at the value of y which minimizes the quantity in curly brackets. For example, if $\rho(|x-y|) = |x-y|$, the respondent should concentrate r(y)dy at the median of his subjective distribution; if $\rho(|x-y|) = |x-y|^2$ he should concentrate its mass at the mean. So this is another scoring scheme which tends to favor the incautious forecaster. In fact, example (a) above may be viewed as a special case of this example, with

$$\begin{aligned} \rho(|i-j|) &= 0 \qquad \text{if } i = j \\ &= 1 \qquad \text{if } i \neq j. \end{aligned}$$

This scoring procedure was not, in fact, used in the Dalkey experiments because its undesirable features were recognized in time. But if it had been used it is easy to imagine some of the erroneous conclusions it would have engendered concerning the relationship between the spread of subjective probability distribution and accuracy (they are related, of course, but not as strongly as this scoring system would have suggested).

(c) A scoring system which we shall call the "Colonel Blotto scoring system" is as follows: suppose you have two or more forecasters making predictions about which of n

mutually exclusive alternatives will take place. Award one point to the forecaster who ascribes the highest probability to the event which actually takes place, and nothing to the others. After a reasonable number of forecasts have been scored in this way, your best forecaster will presumably have accumulated the greatest number of points.

To see what is wrong with this scoring system let us consider the following simple case: you have exactly two forecasters and three equally likely alternatives. Suppose forecaster #1 (correctly) ascribes probability 1/3 to each alternative, while forecaster #2 (through either foolishness or guile) ascribes probability 1/2 to each of the first two alternatives and probability 0 to the third; then the second (inaccurate) forecaster is twice as likely to win the point as is the first (accurate) forecaster. If the first forecaster guesses what kind of ascription the second forecaster is making, he can switch the odds back in his favor by ascribing (.6, .2, .2) to the three possible outcomes, and so on. The two players become locked in the equivalent of a two-person, zero-sum, continuous Colonel Blotto game with one another, and it becomes impossible to infer from their ascriptions what they actually believe are the relative probabilities of the alternatives. This game, by the way, was introduced by Emile Borel in one of the earliest monographs on game theory.[2] He con-

sidered it a very difficult game, but a solution was found thirty years later by Oliver Gross.[11,12] An optimal strategy is the following: label the sides of an equilateral triangle of unit area with the three alternatives under consideration; inscribe a circle in the triangle and erect a hemisphere upon it; choose a point at random on the hemisphere and drop a perpendicular from it to the plane of the triangle; ascribe to each alternative a probability equal to the area of the triangle determined by the foot of the perpendicular and the side corresponding to the alternative in question. Performing this ritual will protect you from being outguessed by your opponent. But the probability you ascribe to a given alternative has little to do with your image of the true probability: In the case we have been discussing it is equally likely to be any number between 0 and 2/3.

3. CLASSES OF REPRODUCING SCORING SYSTEMS

3.1 Preliminaries

In principle, reproducing scoring systems are not necessarily symmetric. That is, if you assign the same probabilities to alternative #1 and alternative #2, your payoff if #1 actually occurs is not necessarily the same as it would be if #2 actually occurs. One could conceive of applications in which this asymmetry would be useful, but in general it seems more reasonable to restrict our attention to symmetric scoring schemes, in which the pay-off depends only on the probability assigned to the alternative which takes place, rather than on the label attached to that alternative. This restriction will also make our mathematical symbolism considerably simpler. The modifications which would be required to extend our results to asymmetric scoring systems will, in most instances, be rather obvious.

Adding a constant factor to a reproducing scoring system yields another reproducing scoring system. It is convenient to normalize a scoring system (to be applied to n alternatives) in such a way that the pay-off for assigning a probability of $\frac{1}{n}$ to the alternative which actually takes place is zero. This makes a certain amount of heuristic sense: If confronted with n alternatives about which you know absolutely nothing, the "principle of equal ignorance" suggests that you should assign equal probabilities to them

all. The reward for this "prophecy" should reasonably be zero.

In what follows, therefore, we shall assume (unless the contrary is specifically stated) that any reproducing serving system we mention is symmetric, and normalized so that the payoff for assigning equal probabilities to all alternatives is zero.

3.2 Two Alternatives

Let us suppose we have two distinct alternatives, one of which must occur and both of which cannot occur. The expert whom we consult reports to us probability r_1 for the first alternative and r_2 for the second. Logic compels him to assign these probabilities in such a way that $r_1 + r_2 = 1$. We agree to pay him $f(r_1)$ if the first alternative comes true, and $f(r_2) = f(1-r_1)$ if the second comes true (a negative payment, as usual, corresponds to his paying us). Let p_1 and p_2 represent the probabilities which, in his heart, our expert actually ascribes to the two alternatives; p_1 and p_2 may not be the same as the probabilities which he reports to us. He will perceive his expected gain from the exercise to be

$$G = p_1 f(r_1) + (1-p_1) f(1-r_1)$$

The essence of a reproducing scoring system is that the expert should perceive his expected gain to be maxi-

mized if he reports to us the probabilities which he actually ascribes to the events in question. If f is differentiable,* a necessary condition for this is the following (in which we suppress subscripts for notational convenience):

$$(3.2.1) \qquad rf'(r) - (1-r)f'(1-r) = 0 \qquad 0 < r < 1$$

Now define $\varphi(r)$ as follows:

$$rf'(r) = \varphi(r) \qquad 0 < r < 1$$

Since $\varphi(1-r) - \varphi(r) = (1-r)f'(1-r) - rf'(r) = 0$, we may also write

$$rf'(r) = \varphi(r) + r[\varphi(1-r) - \varphi(r)]$$

and thus, since $f\left(\frac{1}{2}\right) = 0$,

$$(3.2.2)\ f(r) = \int_{\frac{1}{2}}^{r} \frac{\varphi(t)}{t}\, dt - \left[\int_{\frac{1}{2}}^{r} \varphi(t)dt + \int_{\frac{1}{2}}^{1-r} \varphi(t)dt\right]$$

We have shown that any differentiable reproducing scoring function f(r) for two alternatives may be put in the form of Eq. (3.2.2). In the next section we shall show

*It is proved in App. A that f is monotone increasing; thus it must be differentiable almost everywhere.

that if $\varphi(x)$ is positive except on a set of measure zero, any function defined as in Eq. (3.2.2) is a reproducing scoring function.

3.3 The Gambling House Construction Method

Suppose our expert enters a gambling house in which there are an infinite number of altruistic gamblers. Each gambler corresponds to a point on the interval from zero to one; the gambler corresponding to x is willing to accept an infinitesimal positive wager $\varphi(x)dx$ that alternative #1 will occur, and offers odds of one for x. That is to say, if our expert loses the wager he loses $\varphi(x)dx$, while if he wins the wager he wins $\frac{\varphi(x)}{x}\,dx - \varphi(x)dx$.

Obviously our expert will perceive it to be in his interest to place bets with those gamblers (and only with those gamblers) who correspond to $x \leq p_1$; for he will feel that all these wagers are at favorable or fair odds, while any bets with gamblers corresponding to $x > p_1$ would be placed at odds which appear unfavorable to our expert. If he then visits a second gambling house which is exactly like the first except that the gamblers in the second house accept wagers on the occurrence of alternative #2, then it is easy to see that his net gain, if alternative #1 occurs, will be

$$(3.3.1)\quad g(p_1) = \underbrace{\int_0^{p_1} \frac{\varphi(x)dx}{x}}_{\text{Payments from gamblers in first house}} - \underbrace{\int_0^{p_1} \varphi(x)dx}_{\text{Payments to gamblers in first house}} - \underbrace{\int_0^{1-p_1} \varphi(x)dx}_{\text{Payments to gamblers in second house}}$$

This defines a symmetric reproducing scoring system, but it offers better than zero payoff at $p_1 = p_2 = \frac{1}{2}$. To normalize the system we must subtract a constant; this is equivalent to making the lower limits on all the integrals $\frac{1}{2}$ instead of 0. Thus we are left with Eq. (3.2.2). In other words, Eq. (3.2.2) "characterizes" symmetric reproducing scoring systems. This theorem was discovered independently by Shuford, Albert and Massengill[18] and by Aczel and Pfanzagl.[1]

This "gambling house" construction method has the very desirable feature that it generalizes directly to more than two alternatives. The same discussion as we went through above for two alternatives shows that, if $\varphi(x)$ is any positive function whatsoever, then the following is a reproducing scoring system for n alternatives:

$$(3.3.2)\quad f_i(r_1, r_2, \ldots, r_n) = \int_{\frac{1}{n}}^{r_i} \frac{\varphi(x)dx}{x} - \sum_{j=1}^{n} \int_{\frac{1}{n}}^{r_j} \varphi(x)dx$$

When we are dealing with n alternatives and a symmetric scoring system, we may write the scoring function in

the form f(r,v), where r represents the probability ascribed to the alternative which actually occurs and v is some symmetric function of the r_i's (r_i, of course, being the probability ascribed to the i^{th} alternative). Let us look at two specific reproducing scoring systems which are generated by Eq. (3.3.2).

(a) The quadratic scoring system. Let $\varphi(x) = 2x$, and we get

$$f(r,v) = 2r - v - \frac{1}{n}, \quad v = \sum_{i=1}^{n} r_i^2$$

This scoring system was used by Bruno de Finetti[5] in a series of studies involving football forecasts. It may be shown that, in the case n = 2, the quadratic scoring system is the only one in which the difference between the expected pay-off of a "perfect" expert and of a given expert is a function only of the difference between the "true" probability and the probability ascribed by the given expert.*

(b) The logarithmic scoring system. Let $\varphi(x) = 1$, and we get

$$f(r,v) = \log(n \cdot r)$$

We shall see in the next section that, for n > 2,

*See Appendix C.

this is essentially the only differentiable reproducing scoring system which has the property that the payoff depends only on the probability ascribed to the event which actually occurs (and not on the probabilities ascribed to events which do not occur). In other words, it is the only f(r,v) which does not depend explicitly on v. This fact was evidently first observed by A. H. Gleason (see 15), but the only published proofs I have found are by Aczel and Pfanzagl[1] and Shuford, Albert, and Massengill.[18]

This scoring system relates in an interesting way to information theory. An expert who reports a spectrum of probabilities $(r_1, r_2, \ldots, r_n)$ will assess his expected payoff as

$$\sum_{i=1}^{n} r_i \log(n \cdot r_i)$$

which is essentially a constant (logn) minus the entropy of his partition. In other words, his expected payoff is exactly the same as the amount by which he is able to reduce the expected information in the event itself.

3.4 Many Alternatives

We demonstrated in Secs. 2 and 3 above that all differentiable two-alternative reproducing scoring schemes are generated by the "gambling house" method (Eq. (3.3.2)). The following example shows that this may not be

true for n-alternative schemes:

$$f(r,v) = \frac{r}{\sqrt{v}} - \frac{1}{\sqrt{n}}, \qquad v = \sum_{i=1}^{n} r_i^2$$

This is the so-called "spherical" scoring scheme which was invented by Masanao Toda.[22] If n = 2 it is generated by

$$\varphi(x) = \frac{x}{2[x^2 + (1-x)^2]^{3/2}} .$$

For $n > 2$, however, I have been unable to find a $\varphi(x)$ which will, when plugged into Eq. (3.3.2), give the required function.

Although we are not able to give a characterization of differentiable n-alternative reproducing scoring systems such as we gave for 2-alternative systems, we are able to derive a useful necessary condition that such systems must fulfill, analogous to Eq. (3.2.1). By taking appropriate directional derivatives we are able to determine the following set of n equations which f(r,v) must satisfy:

$$3.4.1) \quad r_i \left.\frac{\partial f}{\partial r}\right]_{r=r_i} + \frac{\partial v}{\partial r_i}\left\{\sum_{j=1}^{n} r_j \left.\frac{\partial f}{\partial v}\right]_{r=r_j}\right\} = w \qquad i = 1, 2, \ldots, n$$

where w is a symmetric function of the r's.* From this

*That is, symmetric on the surface $\sum_{i=1}^{n} r_i = 1$.

equation it is possible to deduce Gleason's result that the logarithmic scoring scheme is essentially the only one in which the payoff depends only on the probability assigned to the alternative which actually occurs. For if f(r,v) is a function of r alone then

$$\frac{\partial f}{\partial v} \equiv 0$$

so we have

$$r_i \left[\frac{\partial f}{\partial r}\right]_{r=r_i} = w.$$

But the left-hand side is a function of r_i alone, while the right-hand side is a symmetric function: The only symmetric function (if $n > 2$) which does not depend on the other r's is the constant function. Thus $w = c$ and

$$\frac{\partial f}{\partial r} = \frac{c}{r}$$

From which it follows that f is a logarithmic function of r.

The logarithmic reward function has the peculiarity that, if the expert in question assigns zero probability to an event which actually takes place, the penalty levied against him is infinite. This may or may not seem

inappropriate.* In case it does seem inappropriate, Shuford, et al.,[18] recommend simply truncating the reward function at some small $\epsilon > 0$ and giving the expert the same reward for $r_i < \epsilon$ as he gets for $r_i = \epsilon$. This reward structure is not a reproducing scoring system for small values of r, of course. Another approach which is more consistent with the spirit of reproducing scoring systems is to use, instead of the log generator $\varphi(x) = 1$, the following φ:

$$(3.4.2) \qquad \varphi(t) = tKn \qquad t < \frac{1}{Kn}$$

$$= 1 \qquad \text{elsewhere}$$

K stands for some large constant. This φ leads to the following pay-off function:

$$f(r, v) = \log nr - v \text{ if } \quad r \geq \frac{1}{Kn}$$

$$= Knr - \log K - 1 - v \text{ if } r < \frac{1}{Kn}$$

$$v = \sum_{\substack{\text{all } i \text{ s.t.} \\ r_i < \frac{1}{Kn}}} \frac{[1 - r_i Kn]^2}{2Kn}$$

*Thomas L. Hughes, former head of intelligence for the U.S. State Department, tells the story of a certain British intelligence officer who, upon his retirement in 1950 after forty-seven years of service reminisced: "Year after year the worriers and fretters would come to me with awful predictions of the outbreak of war. I denied it each time. I was only wrong twice."

It is easy to see that if there are no estimates less than 1/Kn, this payoff function is identical with the logarithmic one.

3.5 Continuous Cases

Let us suppose you are trying to elicit from your experts an estimate of what the population of Uganda will be in 1980. One way to proceed would be to ask the experts to estimate the probability that the population would be less than 7 million, between 7 million and 8 million, between 8 million and 9 million and so on. Unfortunately the way in which you structure your alternatives may have an undue influence on the experts' answers, and if you wish to combine an opportunity for precision with a wide range of possible answers the number of alternatives may become impossibly large. A different approach which seems preferable is to ask the experts directly to specify a probability distribution over the possible future population of Uganda in some way (possibly by asking for percentile breaks, for example) and then score this continuous distribution directly. It appears that any n-alternative reproducing scoring system may be converted, by a limiting process as n becomes large, into a scoring system for continuous distributions on the real line. The details of the limiting process will differ from case to case; we will discuss three examples.

(a) Quadratic Scoring System. Let us suppose that the domain of possible answers is of length D, and is

divided into n equal segments, each of length Δx. If $r(x)$ is an estimating probability density function on D, then the quadratic reward function with respect to this partition would be (assuming $r(x)$ is approximately constant over each interval Δx)

$$3.5.1) \quad f(x_i) = 2r(x_i)\Delta x - \sum_j [r(x_j)\Delta x]^2 - \frac{\Delta x}{D}$$

The last term in this expression, the only one which depends explicitly on D, is only a constant intended to normalize the reward function so that assigning equal probability to all alternatives gets reward zero. Since in this case it is convenient to be noncommittal about what is meant by "all alternatives" we will delete this constant (this deletion does not, of course, destroy the reproducing property of the quadratic scoring system). For the reward function specified by Eq. (3.5.1) to remain nonzero as $\Delta x \to 0$, it is necessary to divide out Δx. The resulting sequence of reward functions then approaches, in the limit,

$$(3.5.2) \qquad f(x) = 2r(x) - \int_D [r(t)]^2 \, dt$$

Let us illustrate the application of this scoring system to two estimates of the population of Uganda in

1980. Suppose two experts are asked to estimate the population in millions. Expert number one gives the following distribution:

$$r_1(x) = \frac{1}{5} \qquad 7 < x < 12$$

$$= 0 \qquad \text{elsewhere}$$

Suppose expert number two gives the following distribution:

$$r_2(x) = \frac{3}{5} \qquad 9 < x < 10$$

$$= \frac{1}{5} \qquad 8 < x < 9, \ 10 < x < 11$$

$$= 0 \qquad \text{elsewhere}$$

Then we can calculate their payoffs under various contingencies in the following table:

TRUE VALUE	PAYOFF TO EXPERT #1	PAYOFF TO EXPERT #2
$x < 7$	− .2	− .44
$7 < x < 8$	+ .2	− .44
$8 < x < 9$	+ .2	− .04
$9 < x < 10$	+ .2	+ .76
$10 < x < 11$	+ .2	− .04
$11 < x < 12$	+ .2	− .44
$12 < x$	− .2	− .44

Note that although both experts assigned the same probability to x falling between 8 and 9, expert #1 gets a positive reward if that contingency comes to pass while expert #2 gets a negative reward. The justification for this is, of course, that expert #2 thought that x was much more likely to fall between 9 and 10, and he "dropped a bundle" betting on that contingency.

(b) Spherical Scoring System. Under the assumptions made above, the spherical scheme would give us

$$(3.5.3) \qquad f(x_i) = \frac{r(x_i)\Delta x}{\sqrt{\Sigma r(x_j)^2 \Delta x^2}} - \frac{1}{\sqrt{D/\Delta x}}$$

We ignore the constant term. The Δx's cancel one another, but since the denominator becomes large as $n \to \infty$ we must divide the reward by $\sqrt{\Delta x}$ in order to keep it from approaching zero as $\Delta x \to 0$. We then have in the limit the following scheme:

$$(3.5.4) \qquad f(x) = \frac{r(x)}{\sqrt{\int_D r(t)^2 dt}}$$

Applying Eq. (3.5.4) to the distribution given in the previous example generates the following table:

TRUE VALUE	PAYOFF TO EXPERT #1	PAYOFF TO EXPERT #2
$x < 7$	0	0
$7 < x < 8$	.447	0
$8 < x < 9$	.447	.302
$9 < x < 10$	.447	.904
$10 < x < 11$	.447	.302
$11 < x < 12$	.447	0
$12 < x$	0	0

(c) Logarithmic Scoring System. If the same type of limit operation carried out in the example above is applied to the logarithmic scoring system we come out with merely the following:

$$f(x) = \log r(x) \tag{3.5.5}$$

Applying 3.5.5 to the two distributions discussed in the preceding example leads to the following table.

TRUE VALUE	PAYOFF TO EXPERT #1	PAYOFF TO EXPERT #2
$x < 7$	$-\infty$	$-\infty$
$7 < x < 8$	-1.609	$-\infty$
$8 < x < 9$	-1.609	-1.609
$9 < x < 10$	-1.609	$-.511$
$10 < x < 11$	-1.609	-1.609
$11 < x < 12$	-1.609	$-\infty$
$12 < x$	$-\infty$	$-\infty$

Note that this is the only case in which we get away from the difficulty (if you consider it such) of giving

different payoffs to experts who assess the same probability for the outcome actually occurring.

4. APPLICATIONS

4.1 Introduction

Reproducing scoring systems have been studied, mathematically and in experimental exercises, for about fifteen years now. Only recently, however, have they been used in practical applications. In this section we will review some of these applications and suggest how reproducing scoring systems could be applied to improve political and economic forecasting in Delphi-type procedures.

4.2 Testing Students

All of us have occasionally taken true-false tests in which the instructor claimed to "count points off for guessing." By this he usually meant that he awarded +1 for a right answer and -1 for a wrong answer. Any student soon realizes that this is not really penalizing guessing at all, but rather that he will maximize his expected score by putting down some answer to every question, even if he thinks that the chance his answer is right is only slightly greater than the chance that it is wrong. A student putting down "true" or "false" to a question has no way of indicating his degree of uncertainty as to the correctness of his answer. This in turn means that the teacher who uses objective tests can only get an accurate reading on the state of the class' knowledge by administering rather long tests.

It would seem reasonable to let the students mark a question "true with probability .80" instead of requiring them to say "true with probability 1.00" or "true with probability 0." Indeed, the latter approach may be subtly training students to tend toward extreme opinions. The use of reproducing scoring systems offers a way of inducing them to put down their true subjective feelings about each question, and this should vastly increase the amount of information which the teacher gets about a class from a given set of questions.

This application has been vigorously championed by Shuford and Massengill[20] and they have marketed materials which enable teachers to understand and apply reproducing scoring systems in the classroom without getting involved in difficult computations. These techniques have been applied in the Academic Instructors Course at Air University, and in pilot programs at the Air Force's Chanute Technical Training Center, the U.S. Army Signal Center and School at Fort Monmouth, the Naval Service School Commands at Great Lakes and at San Diego, and the Naval Air Basic Training Command at Pensacola.[17] Besides offering an improved technique for written objective tests, reproducing scoring systems offer a good method for predicting whether a person can or cannot perform a given practical task[19]: You simply ask him what he thinks the probability is that he can successfully perform the task. You <u>may</u> then ask

him to actually do it: If you do, then he is rewarded (or penalized), whether he succeeds or fails, according to the probability of success he predicted. It is obviously not necessary to ask every student to perform every task in order to induce them, in such a system, to give the most realistic self-estimates of which they are capable.

Because these applications are so new, it will probably be some time before it is completely clear whether they are in fact as advantageous as they appear to be in theory. However a test or quiz is a communication device between students and teacher; by modifying objective tests in the way indicated in this section the channel capacity of the communication device is increased and somehow this must be a good thing.

4.3 Weather Forecasting

Meteorologists have, for some time, cast many of their predictions in probabilistic terms, and the appropriateness of this has been well recognized.[13] More recently, reproducing scoring systems (or "proper scoring systems," as the meteorologists call them) have been proposed[10,25] and actually applied[24] to the evaluation of such forecasts.

Reproducing scoring systems may be used, of course, not only to compare the merits of different human forecasts, but also to compare the quality of different mathematical algorithms for preparing probabilistic forecasts, and for comparing such mechanical forecasts with those made by

flesh and blood human beings.[24] In fact, the logarithmic scoring rule, when used to distinguish which of two probabilistic models gives the better fit to the observed data, is identical with the well-known maximum likelihood criterion.[23]

4.4 Delphi Processes

The "Delphi process" is a name which has been given to a technique in which a group of individuals independently estimate some quantity, and arrive at an improved estimate by carefully controlled communication with one another.[3,4] There is no reason why the quantity they are attempting to estimate should not be the probability of some future event.

To be specific, let us envision the following means of forecasting the news events of 197X. A panel of twenty or more knowledgeable individuals is selected, and each is asked to answer, without consulting the others, the same set of questions about probable events during the year. Some typical questions might be the following:

"What is the probability that North Korean forces will destroy or capture a United States intelligence vehicle during the first six months of 197X?"

"What is the probability that at least one U.S. astronaut will be killed in line of duty during 197X?"

"How many Republicans will be elected to the House of Representatives in November?

There is a .01 chance of less than _____
" " " .10 " " " " _____
" " " .30 " " " " _____
" " " .30 " " more " _____
" " " .10 " " " " _____
" " " .01 " " " " _____"

"What is the probability that the number of U.S. troops in Vietnam on October 1 will be

less than 100,000? _____
Between 100,000 and 300,000? _____
" 300,000 " 400,000? _____
" 400,000 " 500,000? _____
" 500,000 " 600,000? _____
Over 600,000? _____."

After the participants had made their various estimates, they might be given, say, the median estimate produced by the group and individually asked if they would like to change their estimate.

It would be explained to the participants that they would each be paid for their efforts at the end of 197X, and that this payment would be in proportion to their individual scores as calculated by an appropriate reproducing scoring system. Enough money should be set aside to pay them fees whose expected value would be adequate to justify a good effort.

The results of such a survey might make an entertaining thirty-minute TV news special; but would they be of real help in solving any real-world problems? Would the forecasts (or the "average" forecast) be more reliable than forecasts derived by other methods? Would the forecasts improve significantly if such a survey was taken year after year, with greater weight being given to the seasoned experts who showed high accuracy in preceding surveys? My inclination is to answer all of these questions in the affirmative, but barring an actual trial there is no way of answering any of them with assurance.

5. POTENTIAL DIFFICULTIES

5.1 Irrationality

The concept of applying reproducing scoring systems to the respondents in a Delphi panel is founded on the notion that individuals generally behave "rationally" in the sense that, under conditions of uncertainty, they will act in such a way that they maximize their expected gains. Unfortunately there is some evidence that this is not the case.[20,21] For example, when confronted with a choice between gambles, some people will choose the one with the highest probability of winning something rather than the one with the highest expected payoff; others will choose the gamble with the highest top prize regardless of how remote the chance of winning it may be. It may be that the sort of person we are apt to select for a Delphi panel is not likely to behave in those "irrational" ways; but if we do have the bad luck to find a large number of such people on our panel then it is obvious that any reproducing scoring system will have highly unpredictable effects. It may provide the panel with incentives to exaggerate or understate their true subjective probabilities, and the resulting reports may thus turn out to be less meaningful than if we had used no scoring system whatsoever.

To guard against such a possibility it would seem wise, if a substantial number of panelists were not well-known to you, to include certain questions on your Delphi

questionnaire whose sole purpose would be to determine if the respondent was "rational" or not. Unfortunately such questions would be, perforce, somewhat artificial—such as "What is the probability that Army will kickoff at the beginning of the Army-Navy game next fall?"

5.2 Unintended Payoffs

It is possible for a panel to be entirely "rational" in the sense of the preceding section, and still give distorted responses simply because the "expected gains" which they are maximizing include factors which you have neglected to consider. For example, the panelists might have special interest in having you adopt a particular point of view or course of action.

Indeed, the world of forecasting is today awash with conflict of interest at all levels. Your family physician predicts that if you let him perform a $3,000 operation you will never be troubled with leg cramps again; your lawyer predicts victory in litigation while collecting his retainer; an Air Force general predicts victory in Vietnam if we step up the bombing; an electronics firm predicts marvelous bombing accuracies in five years if we spend more on research and development; many research reports (including this one) implicitly or explicitly predict high payoffs from further research. The self-serving nature of many prophecies is often obvious, but it is more dangerous when it is subtle and well hidden.

It would be attractive to suppose that using a reproducing scoring system and making rewards proportional to score would do away with the problem of conflict of interest. Of course it would not. The president of an electronics firm interested in building a new bomb-nav system is going to predict high confidence in high performance for the new system regardless of what kind of explicit scoring system you apply to his prophecies, simply because he has so much more to gain from getting the contract than he has to gain from any payments you might realistically make for accurate forecasting. Reproducing scoring systems do make it possible to bring in whole new groups of forecasters, individuals who ordinarily would not bother to work out a forecast in a given area because it does not directly affect their personal interests. Such people could be influenced by a properly designed system to work out forecasts in which the compulsion of personal interest was all toward as much accuracy as they are capable of. But it would be a mistake to suppose that holding out one of the scoring systems discussed in this paper to the same people who are now producing self-serving forecasts would suddenly induce them to produce unbiased forecasts in the future.

Another possible unintended payoff is similar to the "Colonel Blotto" scoring system we discussed previously.

Suppose you set up a Delphi experiment with five

questions and a thousand participants, and announced that you would use, say, the quadratic scoring system and would award $1,000 to whoever achieved the top score. This does not constitute a reproducing scoring system. For suppose each question was a two alternative forecast with true probability of .50: Any participant who perceived the true probabilities and reported them would, with certainty, achieve a score of zero. On the other hand, a crafty speculator who expresses "certainty" on all five questions has one chance in thirty-two of making an unbeatable score of 2.5. His expected score is -4, but making a big negative score leaves him no worse off than the cautious clunk who scores zero; the only thing that matters is his chance of coming in first. To calculate the true optimal policy in this system is a very complex matter, since it is essentially a 1,000-person game we are discussing, but it is clearly not wise to make "honest" forecasts.

In actual cases, of course, the pressure toward extreme forecasts caused by this top-dog syndrome is apt to be considerably more subtle. For example, one of the functions we have repeatedly stressed for scoring systems is distinguishing good forecasters from poor ones. If the members of your panel get the impression that the top forecasters will get special recognition (more than just a monetary award proportional to their score) this will

introduce a distorting tendency into their forecasts.

Finally there is the well-known nonlinear utility of money. Twenty thousand dollars is "worth" less than twice as much as ten thousand dollars. The conflict of interest problem and the top-dog syndrome argue in favor of making the monetary award for accurate forecasting substantial: But if you make the awards too substantial you may find yourself in a situation where some of your respondents become overcautious, and prefer to go for small but relatively certain gains in preference to maximizing their expected gain through a riskier strategy.

5.3 Slow Discrimination

Reproducing scoring systems can indeed be used to determine which of two probabilistic forecasters is the more accurate, but it may take many trials before you can place much confidence in this determination.

For example, suppose two experts are asked to predict the probability of occurrence of n different events, and each event has true probability one-half. Suppose expert number one ascribes probability three-fifths to each event. Suppose we are using a quadratic reproducing scoring system. Let d(n) denote the difference between the first expert's score and the second's (actually the first expert will, if the scoring system is properly normalized, make zero score, so -d(n) will be the second expert's score). The quantity d(n) is a random variable; it may be either

positive or negative, but its mean is positive and the ratio of its mean to its standard deviation may be calculated to be $\sqrt{n}/10$. This means that even after 100 predictions have been examined, there is still about one chance in six that the second (less accurate) forecaster will outscore the first. After 400 predictions the chance of this happening is down to one in forty. A great many forecasts, in other words, have to be evaluated before you can place much confidence in the leading scorer being truly the best forecaster.

In fact, it can be shown* that under any reproducing scoring system whatsoever, if d(n) represents the difference after n trials between the score of an expert who ascribed probability y_2 and one who ascribed probability y_1 to each event (and the true probability is p), then

$$\sqrt{\frac{n}{p(1-p)}}\,[p - y_1] < \frac{\text{mean } (d(n))}{\text{stand. dev. } (d(n))} < \sqrt{\frac{n}{p(1-p)}}\,[p - y_2].$$

From this we see that experts will be discriminated more quickly if they are asked to forecast high probability events than middling probability events, but that in any case a fairly large number of forecasts are apt to be required.

---

*See Appendix D.

Appendix A

MONOTONICITY

Let f be a reproducing scoring system on two variables. Define

$$H(x, p) = pf_1(x) + (1-p)\ f_0(1-x)$$

H(x, p) is clearly the expected payoff of an expert assessing the probability of an event as x when its true probability is p.

THEOREM A: If $p \leq x < y$ or $y < x \leq p$, then

$$H(x, p) > H(y, p).$$

PROOF (Due to John Lindsey): By definition,

$$(A.1)\quad xf_1(x) + (1-x)\ f_0(x) > xf_1(y) + (1-x)\ f_0(y)$$

$$(A.2)\quad yf_1(y) + (1-y)\ f_0(y) > yf_1(x) + (1-y)\ f_0(x)$$

Now assume $p \leq x < y$. Then

$$\alpha = \frac{y-p}{y-x} > 0$$

$$\beta = \frac{x-p}{y-x} \geq 0$$

Multiplying A.1 by α, A.2 by β, and adding, gives us

$$(A.3) \quad pf_1(x) + (1-p)\, f_0(x) > pf_1(y) + (1-p)\, f_0(y).$$

Which was to be proved. The same argument works for $y < x \leq p$.

COROLLARY: $f_1(x)$ is a monotone increasing function of x. $f_0(x)$ is a monotone decreasing function of x.

PROOF: $f_1(x) = H(x, 1)$. $f_0(x) = H(x, 0)$.

Appendix B

NONUNIQUENESS OF φ

In Sec. 3 we showed that any symmetric differentiable reproducing scoring system on two alternatives f could be put in the form

$$\text{B.1} \qquad f(r) = \int_{\frac{1}{2}}^{r} \frac{\varphi(t)}{t}\, dt - \left[\int_{\frac{1}{2}}^{r} \varphi(t)dt + \int_{\frac{1}{2}}^{1-r} \varphi(t)dt\right]$$

where $\varphi(1-t) = \varphi(t)$ (i.e., φ is symmetric about $\frac{1}{2}$). On the other hand, we also showed that any function of the form B.1 is a symmetric reproducing scoring system if φ is positive, whether φ is symmetric about $\frac{1}{2}$ or not. The explanation for this seeming contradiction is that quite different φ may give rise to the same f.

Direct computation shows that, if g(t) is any integrable function, then

$$\text{B.2} \qquad \psi(t) = t\left[\varphi(1-t) - \varphi(t)\right]$$

is a solution to the integral equation

$$\text{B.3} \qquad 0 = \int_{\frac{1}{2}}^{r} \frac{\psi(t)}{t}\, dt - \int_{\frac{1}{2}}^{r} \psi(t)dt - \int_{\frac{1}{2}}^{1-r} \psi(t)dt$$

This is the most general solution, since if $\psi(t)$ is any solution differentiation of B.3 shows that

B.4 $$\psi(t) = t[\psi(t) - \psi(1-t)]$$

Now let $\varphi(t)$ be any positive, integrable function defined over $[0,1]$. Then

B.5 $$\varphi_1(t) = \varphi(t) + t[\varphi(1-t) - \varphi(t)]$$

is positive and symmetric and when plugged into B.1 gives exactly the same f as does φ.

Appendix C

AN INVARIANCE PROPERTY CHARACTERIZING THE QUADRATIC SCORING SYSTEM

It would seem desirable to find a symmetric reproducing scoring system on two alternatives with the property that an individual's expected reward would depend only on his accuracy, and not upon the value of the true probability of the event in question. A little reflection shows that there can be no reproducing scoring system with this property: We may assume the system normalized so that the payoff for perfect accuracy when $p_1 = p_2 = \frac{1}{2}$ is zero. Then the payoff for perfect accuracy when $p_1 = 1$, $p_2 = 0$ must also be zero, so $f(\frac{1}{2}) = f(1) = 0$ which contradicts the theorem of Appendix A (that f is monotone increasing).

Even though it is impossible to find a nontrivial reproducing scoring system which makes an individual's expected reward independent of p_1 it is possible to find a scheme which makes an individual's relative expected reward (i.e., the difference between his reward and that expected by a hypothetical perfect expert) depend only on p-r and not on p. That is to say, the difference in expected rewards is only a function of the difference between the individual's forecast and the true probability. To be more specific, let us define E(r,p) to be the expected return of an individual who ascribes probability r to an event whose true probability is p. That is,

C.1 $$E(r,p) = pf(r) + (1-p)f(1-r)$$

Now consider the quadratic reproducing scoring system

C.2 $$f(r) = 4r - 2r^2 - \frac{3}{2} = \frac{1}{2} - 2(1-r)^2$$

Simple calculation shows that

C.3 $$E(p,p) - E(r,p) = 2(p-r)^2$$

Thus the quadratic scoring system has the property we desire; we will now demonstrate that it is essentially the only scoring system which does.

THEOREM: Let f(r) be a differentiable symmetric reproducing scoring system on two alternatives which satisfies the following conditions:

(1) $E(p,p) - E(r,p)$ is a function of $(p-r)$.
(2) $f(\frac{1}{2}) = 0$.
(3) $f(1) = 1$.

Then

$$f(r) = 1 - 4(1-r)^2$$

PROOF: Let

C.4 $$E(p,p) - E(r,p) = h(p-r).$$

By Eq. (2) we see that $E(\frac{1}{2},p) = 0$, so

C.5 $$E(p,p) = h(p-\tfrac{1}{2})$$

Combining C.4 and C.5 we have

C.6 $$h(p-\tfrac{1}{2}) - h(p-r) = E(r,p) = pf(r) + (1-p)f(1-r)$$

Let $r = \frac{1}{2} + \epsilon$, and divide both sides of C.6 by ϵ, and we find

C.7 $$\frac{h(p-\frac{1}{2}) - h(p-\frac{1}{2}-\epsilon)}{\epsilon} = \frac{f(\frac{1}{2}-\epsilon) - f(\frac{1}{2})}{\epsilon} + p\left[\frac{f(\frac{1}{2}+\epsilon) - f(\frac{1}{2}-\epsilon)}{\epsilon}\right]$$

Let $\epsilon \to 0$. The limit on the right-hand side exists since f is differentiable; thus the limit on the left-hand side exists and we have

C.8 $$h'(p-\tfrac{1}{2}) = f'(\tfrac{1}{2})\ (2p-1).$$

We know $h(0) = 0$, and thus solving the differential equation C.8 gives us

C.9 $$h(p-\tfrac{1}{2}) = f'(\tfrac{1}{2})\ (p-\tfrac{1}{2})^2$$

since

C.10 $\quad h(\frac{1}{2}) = E(1,1) - E(\frac{1}{2},1) = E(1,1) = f(1) = 1$

it follows that $f'(\frac{1}{2}) = 4$. Thus

C.11 $\quad h(y) = 4y^2$

since

C.12 $\quad h(1-x) = E(1,1) - E(x,1) = 1 - f(x)$

we see at once that

C.13 $\quad f(x) = 1 - 4(1-x)^2$

which concludes the proof. Note that we did not use f's differentiability (or even continuity) except at the point $x = \frac{1}{2}$.

Appendix D

BOUNDS ON DISCRIMINATION

One purpose of a reproducing scoring system is to attempt to single out which experts (of a group) are the most accurate in their estimates of the probability of given events taking place. The rating system itself may, of course, give ratings which are faulty due to "bad luck." For example: if expert #1 predicts that a given fair coin will come up heads on its next toss with probability .50, while expert #2, not understanding that it is a fair coin, fairly flipped, makes a prediction that it will certainly come up heads, and the coin does come up heads, then any of the rating schemes we have been discussing will identify expert #2 as the more accurate of the two. Over a long series of flips, of course, expert #1 will expect to surpass the total rating of expert #2 (providing the latter continues to predict heads with certainty). In this section we will discuss the probability that a reproducing scoring system will fail (that is, that it will rate an inaccurate expert more highly than an accurate one) over a given set of predictions.

In general this probability will be a complex function of the spectrum of true probabilities and estimates encompassed in the set of predictions under consideration. For simplicity, we shall limit our consideration to two-alternative symmetric scoring systems, and assume that

expert #1 and expert #2 estimate that the probability of an event taking place as y_1 and y_2, respectively, and that the true probability of the event is p. If this situation recurs n times, what is the probability that expert #1 will have a higher total score than expert #2? Let d(n) denote the difference between the total score of expert #1 and the total score of expert #2 after n "trials." Of course d(n) is a random variable. If it is positive, then #1 outscores #2; if it is negative, then #2 outscores #1. It is easy to calculate that

$$\text{(D.1)} \quad \text{mean } (d(n)) = n\,[p\{f(y_1) - f(y_2)\} + (1-p)\{f(1-y_1) - f(1-y_2)\}]$$

$$= n\{f(1-y_1) - f(1-y_2)\}$$

$$+ np\{f(y_1) - f(y_2) - f(1-y_1) + f(1-y_2)\}$$

$$\text{(D.2)} \quad \text{variance } (d(n)) =$$

$$np(1-p)\{f(y_1) - f(y_2) - f(1-y_1) + f(1-y_2)\}^2$$

Let us assume that $y_1 > y_2$, so that

$$\text{(D.3)} \quad \{f(y_1) - f(y_2) - f(1-y_1) + f(1-y_2)\} > 0.$$

Then, we have

(D.4) $$\frac{\text{mean } (d(n))}{\text{stand. dev. } (d(n))} =$$

$$\sqrt{\frac{n}{p(1-p)}} \quad \frac{1}{\dfrac{f(y_1) - f(y_2)}{f(1-y_1) - f(1-y_2)} - 1} + \sqrt{\frac{np}{1-p}}$$

Let φ be the symmetric function which (in equation B.1) defines f. Then

(D.5)
$$\frac{f(y_1) - f(y_2)}{f(1-y_1) - f(1-y_2)} = \frac{\int_{y_2}^{y_1} \frac{\varphi(t)dt}{t}}{\int_{1-y_2}^{1-y_1} \frac{\varphi(t)dt}{t}}$$

$$= \frac{\int_{y_2}^{y_1} \frac{\varphi(t)dt}{t}}{\int_{y_2}^{y_1} \frac{\varphi(t)dt}{1-t}}$$

By the mean value theorem

(D.6) $$\frac{1-y_1}{y_1} \int_{y_2}^{y_1} \frac{\varphi(t)dt}{1-t} < \int_{y_2}^{y_1} \frac{\varphi(t)dt}{t} < \frac{1-y_2}{y_2} \int_{y_2}^{y_1} \frac{\varphi(t)dt}{1-t}$$

Thus

$$(D.7) \qquad \frac{1 - y_2}{-y_2} < \frac{f(y_1) - f(y_2)}{f(1-y_1) - f(1-y_2)} < \frac{1 - y_1}{-y_1}.$$

Plugging inequality (D.7) into Eq. (D.4) gives us

$$(D.8) \qquad \sqrt{\frac{n}{p(1-p)}}\ [p-y_1] < \frac{\text{mean } (d(n))}{\text{stand. dev. } (d(n))} < \sqrt{\frac{n}{p(1-p)}}\ [p-y_2].$$

This inequality is useful because it does not contain the particular reward function f. The absolute value of the ratio of the mean to the standard deviation measures how likely the rating scheme is to make a misrating. If n is large enough that d(n) is approximately normal, then the probability of misrating will be less than .025 if the absolute value of the mean over the standard deviation is greater than 2. One might imagine that it would be possible to choose f so cleverly that this ratio would be large even if n were not very great. Inequality (D.8) shows that no matter what reproducing scoring system you choose, the ratio will fall between certain limits. Looking at Eq. (D.6) shows that an f which comes close to either limit is dependent on the particular y_1 and y_2 involved, so it appears that there will be no one scheme which is the best discriminator.

We now summarize the results above by stating them as a theorem.

THEOREM: Let f be a symmetric, two-alternative, differentiable, reproducing scoring system. Let d(n) represent the difference in scores after n trials between an expert who predicts y_1 and an expert who predicts y_2 for an event whose true probability is p, where $y_2 < y_1$; then

$$\sqrt{\frac{n}{p(1-p)}}\ [p-y_1] < \frac{\text{mean } (d(n))}{\text{stand. dev. } (d(n))} < \sqrt{\frac{n}{p(1-1)}}\ [p-y_2].$$

From Eq. (D.4) we can calculate the ratio for $f(r) = 1 - 4(1-r)^2$ (the quadratic scoring system). It is

$$\frac{\text{mean } (d(n))}{\text{stand. dev. } (d(n))} = \sqrt{\frac{n}{p(1-p)}}\ \left[p - \frac{y_1 + y_2}{2}\right]. \tag{D.9}$$

Note that this is exactly half way between our theoretical upper and lower bounds.

BIBLIOGRAPHY

1. Aczel, J., and J. Pfanzagl, "Remarks on the Measurement of Subjective Probability and Information," Metrika, Vol. 11, No. 2, 1966, pp. 91-105.

2. Borel, E., "La théorie du jeu et les équations intégrales à noyau symétrique," Comptes Rendus de l'Académie des Sciences, Vol. 173, 1921, pp. 1304-1308. Translation by L. J. Savage in Econometrica, Vol. 21, No. 1, 1953, pp. 97-100.

3. Dalkey, N., "Analyses from a Group Opinion Study," Futures, Vol. 1, No. 6, December 1969, pp. 541-551.

4. Dalkey, N., "An Experimental Study of Group Opinion: The Delphi Method," Futures, Vol. 1, No. 5, September 1969, pp. 408-426.

5. De Finetti, B., "Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item," British Journal of Mathematical and Statistical Psychology, Vol. 18, Part 1, May 1965, pp. 87-123.

6. Di Paola, R. A., "Random Sets in Subrecursive Hierarchies," Journal Ass. Comp. Mach., Vol. 16, No. 4, October 1969, pp. 621-630.

7. Edwards, W., Nonconservative Probabilistic Information Processing Systems, ESD-TR-66-404, Decision Sciences Laboratory, Electronic Systems Division, L. G. Hanscom Field, Bedford, Massachusetts, December 1966.

8. Edwards, W., and L. D. Phillips, "Man as Transducer for Probabilities in Bayesian Command and Control Systems," in G. L. Bryan and M. W. Shelly (eds.) Human Judgements and Optimality, Wiley, 1964.

9. Eisenberg, E., and G. Gale, "Consensus of Subjective Probabilities: The Pari-Mutuel Method," Annals of Math. Stat., Vol. 30, No. 1, March 1959, pp. 165-168.

10. Epstein, E. S., "A Scoring System for Probability Forecasts of Ranked Categories," Journal of Applied Meteorology, Vol. 8, No. 6, December 1969, pp. 985-987.

11. Gross, O., The Symmetric Blotto Game, RM-424, The Rand Corporation, Santa Monica, 1950.

12. Gross, O., and R. Wagner, A Continuous Colonel Blotto Game, RM-408, The Rand Corporation, Santa Monica, 1950.

13. Malone, T. F., "Applied Meteorology," Meteorological Research Reviews: Summaries of Progress from 1951 to 1955, Meteor. Monogr. 3, Nos. 12-20, pp. 152-159.

14. Martin-Löf, P., "The Literature on von Mises' Kollectirs Revisited," Theoria, Vol. 35, 1969, pp. 12-37.

15. McCarthy, J., "Measures of the Value of Information," Proc. Nat. Acad. Sci., Vol. 42, 1956, pp. 654-655.

16. Norvig, T., "Consensus of Subjective Probabilities: A Convergence Theorem," Annals of Math. Stat., Vol. 38, No. 1, February 1967, pp. 221-225.

17. Shuford, E. H., Confidence Testing: A New Tool for Measurement, The Shuford-Massengill Corporation, Lexington, Massachusetts, 1969.

18. Shuford, E. H., A. Albert, and H. E. Massengill, "Admissible Probability Measurement Procedures," Psychometrika, Vol. 31, No. 2, June 1966, pp. 125-145.

19. Shuford, E. H., and D. L. Gibson, A New Method for Predicting Performance, The Shuford-Massengill Corporation, Lexington, Massachusetts, 1969.

20. Shuford, E. H., and H. E. Massengill, How To Shorten a Test and Increase Its Reliability and Validity, The Shuford-Massengill Corporation, Lexington, Massachusetts, 1967.

21. Slovic, P., and J. Lichtenstein, "Importance of Variance Preferences in Gambling Decisions," Journal of Experimental Psychology, Vol. 78, No. 4, 1968, pp. 646-654.

22. Toda, M. Measurement of Subjective Probability Distributions, ESD-TDR-63-407, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Massachusetts, 1963.

23. Winkler, R. L., "Scoring Rules and the Evaluation of Probability Assessors," Journal of the American Statistical Association, Vol. 64, September 1969, pp. 1073-1078.

24. Winkler, R. L., and A. H. Murphy, "Evaluation of Subjective Precipitation Probability Forecasts,"

Proceedings of the First National Conference on Statistical Meteorology, American Meteorological Society, Boston, 1968, pp. 148–157.

25. Winkler, R. L., and A. H. Murphy, "'Good' Probability Assessors," Journal of Applied Meteorology, Vol. 7, No. 5, October 1968, pp. 751–758.