AN EFFECTIVE SCORING RULE FOR PROBABILITY DISTRIBUTIONS

by

Daniel Friedman
Department of Economics
UCLA

UCLA Department of Economics
Discussion Paper #164

November, 1979

ACKNOWLEDGEMENTS

Without relinquishing credit or blame for what follows, the author would like to thank Ray Peters for suggesting the problem which led to this paper, Ted Groves for leading him to the relevant literature, and Joe Ostroy and Ed Leamer for helpful discussions as the work crystallized into its present form.

AN EFFECTIVE SCORING RULE FOR PROBABILITY DISTRIBUTIONS

I. Introduction

The work presented in this paper was originally motivated by the following problem in communications. Suppose one or more experts have knowledge not generally available about some random variable, observable only *ex post*, and one or more decision makers must make *ex ante* choices whose outcome depends in part on the random variable. Further, suppose that the choices and outcomes are sufficiently complicated that more information about the random variable than just its expected value is required, but the decision makers have limits to their ability and desire to absorb detail. In such a situation, it may be reasonable for the experts to convey their knowledge by means of simplified probability distributions. The decision makers will then need some means of evaluating after the fact the quality of the information they received. They also need assurance that it is in the best interest of the experts to direct their own efforts into producing the most accurate[1] probability distribution possible given limitations of time, money and format.

The problem, then, is chiefly one of eliciting personal probability distributions. There is by now a rather substantial literature on elicitation; Savage (1971) in many ways sets the tone. That article emphasizes the use of *scoring rules* to encourage accurate elicitation, the behavioral assumption being that an expert attempts to maximize his expected score. The scoring

[1] In the absence of a well-chosen incentive structure, the experts may indulge in game playing which distorts their stated probability distributions. For instance, casual observation of economic forecasters suggests that experts who feel they have a reputation to protect will tend to produce a forecast near the consensus, and experts who feel they have a reputation to build will tend to overstate the probabilities of events they feel are understated in consensus.

rule, in turn, can be any function of the elicited information and the realization of the random variable ex post, but of course cannot depend on unobservables such as the subjective (or "true") ex ante distribution. Savage concentrates on proper scoring rules; that is, rules that give a maximum expected score to a perfect elicitation. He also introduces, but does not name, the monotonicity property that the better the elicitation, the higher the expected score. Scoring rules with such a monotonicity property, hereafter called effective scoring rules, are particularly appropriate to the communication problem sketched above, because the experts, constrained to simplicity, may not be able to communicate precisely their true probability distributions.

Unfortunately, Savage's methods are designed to elicit only the mean of a probability distribution. Matheson and Winkler (1976) extend many of Savage's results to our problem of eliciting an entire personal probability distribution, but they deal basically with proper scoring rules, not all of which will be effective.

The next section will provide more precise definitions of these ideas as well as some examples. Section III will demonstrate the effectiveness of two well-known scoring rules and discuss their relative merits. The following section is intended to highlight the main ideas by means of a practical application, and the last section touches on some related questions that may be of interest. Appendix A contains some technical notes.

The notation follows Savage where possible and generally accepted mathematical practice otherwise. Rudin (1966) is a handy reference for Section III.

II. Scoring Rules

Let X be a random variable with distribution F. For the most part, we will think of F as the (expert's) subjective probability distribtuion, and assume that it has some density function f.

Let us first consider the case of eliciting only EX, the expectation of X. In this context, a scoring rule is a real valued function s(y,x), where y is the value given by the expert and x is the value of X actually observed. The rule s is said to be (strictly) proper if its f-expected value, $E_f s(y) \equiv \int s(x,y)f(x)dx$ is (strictly) maximized at y=EX, the "true" expectation of X. We will say that s is effective if its f-expected value is monotonic in the following sense:

1. $$E_f s(y) > E_f s(z) \Longleftrightarrow |\ y-EX\ | < |\ z-EX\ |,$$

that is, the expected score is higher, the closer in absolute value is the elicited value to the "true" (but unobservable) expected value. Clearly s will be strictly proper if it is effective. An example of an effective scoring rule is $s(y,x) = -(y-x)^2$.

Modifications of these definitions are required if we wish to elicit the entire density function f of X rather than just the expectation. In this context a scoring rule is a real valued functional S(g,x) defined for all g in some set D of density functions.[2] S is (strictly) proper if its f-expected value, $E_f S(g) \equiv \int S(g,x)f(x)dx$, is (strictly) maximized on D at g=f. It is effective if

[2] D represents, in the communications problem discussed in the introduction, the set of admissible (simplified) distributions from which the expert selects his message. If $f \notin D$, then no proper scoring rule is possible.

2. $E_f S(g) > E_f S(h) \iff d(f,g) < d(f,h)$,

where d is a metric (i.e., distance function) on D. In words, the expected score is a monotone decreasing function of the distance between the true and elicited distributions. Note that if S is effective, and if f ε D, then S is also strictly proper. In this definition, some metric d on D is assumed. The set of effective scoring rules would in general be different if some other metric were used, so the choice of metric is important.

Some examples should help clarify ideas.

a. The naive scoring rule $S(g,x) = g(x)$ simply uses the height of the density function at the realized outcome as the score. Despite its intuitive appeal, this rule is not proper, and and therefore not effective, as can be seen in the following example. Let

$$f(x) = \begin{cases} 1/4 \text{ if } 0 \leq x < 1 \\ 3/4 \text{ if } 1 \leq x < 2 \\ 0 \text{ otherwise;} \end{cases} \qquad g(x) = \begin{cases} 1 \text{ if } 1 \leq x < 2 \\ 0 \text{ otherwise.} \end{cases}$$

Then $E_f S(f) = \int_0^2 f^2(x)dx = 5/8$, but $E_f S(g) = \int_1^2 f(x)dx = 3/4$.

b. Scoring rules are commonly derived from a decision maker's loss function. For instance, if the loss function is $f(a,x) = (a-x)^2$, where a is the "action" taken (i.e., the value assigned to the random variable) and x is the ex post value, then it is easy to see that one minimizes g-expected loss by setting $a = \bar{g}$, the mean of the elicited distribution. Thus, an appropriate scoring rule in this case would be $S(g,x) = -(\bar{g}-x)^2$. Such a rule is proper but not strictly proper, since any g whose mean coincides with the "true" mean $\bar{f}$ receives the maximum f-expected score. Similar results hold for more complex loss functions, but the topic will not be pursued further here since we are primarily concerned with a communication problem and assume no specific knowledge of decision problems.

c. The logarithmic scoring rule $S(g,x) = \log g(x)$ arises from information theory and can be shown to be strictly proper. However, it gives a very stiff penality to underestimating low probabilities and in fact gives an expected score of $-\infty$ to any g which is zero on a subset of support (f) of positive measure. If support (f) $\equiv \{x|f(x) > 0\}$ is not known *a priori*, practical difficulties arise, which the reader can illustrate by taking f and g as above and choosing various h's "between" f and g. Appendix A indicates that the logarithmic scoring rule probably is ineffective in a very general sense.

d. The quadratic scoring rule $S(g,x) = 2(g(x) - \|g_2^2\|$, where $\|g\|_2^2 = \int |g(x)|^2 dx$, arises as analogue of the simple but effective scoring rule for expectations mentioned near the beginning of this section. This scoring rule is well-known to be strictly proper. As to its effectiveness, consider the densities

$$f(x) = \begin{cases} 10 & \text{if } 0 \leq x < 0.10 \\ 0 & \text{otherwise;} \end{cases} \qquad g(x) = \begin{cases} 10 & \text{if } 0.06 \leq x < 0.16, \\ 0 & \text{otherwise;} \end{cases}$$

and $h(x) = \begin{cases} 0.01 & \text{if } 100 \leq x < 200 \\ 0 & \text{otherwise.} \end{cases}$

According to most intuitive notions of distance, g is much closer to f than is h, but $E_f S(g) = -2.00$ while $E_f S(h) = -0.01$. However, we will soon see that there is a simple metric which makesthis rule effective.

e. The spherical scoring rule $S(g,x) = g(x)/\|g\|_2$ arises as a correction to the naive scoring rule and is effective with respect to a natural metric on virtually any set D of probability distributions, as the next section will show.

III. An Effective Scoring Rule

This section will employ some basic techniques from functional analysis, so a few definitions are in order. Let Ω be some measurable subset of R, perhaps R itself, and let $f: \Omega \to R$. Then the p-norm of f is

$$\|f\|_p = \left\{ \int_\Omega |f|^p dx \right\}^{1/p} \quad 1 \leq p < \infty .$$

The space of p-integrable functions is $L_p = \left\{ f \mid \|f\|_p < \infty \right\}$. The usual metric on L_p is $d_p(f,g) = \|f-g\|_p$. The unit sphere in L_p is $B_p = \{f \in L_p \mid \|f\|_p = 1\}$, and in the case p = 2 we have the inner product $(f,g) = \int_\Omega f(x)g(x)dx$, which will always be finite for f, g $\in L_2$. Finally, let D = {bounded continuous density functions on Ω}, and note that $D \subset B_1 \cap L_2$.

Proposition 1. The quadratic scoring rule Q(g,x) is effective on D with respect to the L_2-metric d_2.

Proof: First note that for any f, g $\in L$, $(d_2(f,g))^2 = (f-g, f-g)$, and $E_f Q(g) = 2(f,g) - (g,g)$. Therefore, for any f $\in L_2$ and g, h $\in$ D,

$$\begin{aligned} d_2(f,g) < d_2(f,h) &\iff (f-g, f-g) < (f-h, f-h) \\ &\iff (f,f) + (g,g) - 2(f,g) < (f,f) + (h,h) - 2(f,h) \\ &\iff 2(f,g) - (g,g) > 2(f,h) - (h,h) \\ &\iff E_f Q(g) > E_f Q(h). \qquad \text{QED.} \end{aligned}$$

The L_2-metric is widely used and is quite natural in the sense that it is a direct extension of ordinary Euclidean distance to an infinite dimensional function space. However, it seems not quite appropriate for use on a space D of density functions; the normalization employed in forming D (integral = 1) does not blend well with the L_2-norm underlying this metric. Thus, in example c of the previous section , the L_2-distance between f and g is exaggerated relative to that between f and h by the fact that the L_2- norms of f and g are much larger than that of h.

Proposition 2. For any $f \in B$, the naive scoring rule $S(g,x) = g(x)$ achieves a unique maximum in f-expected value on B_2 at $g = f$. Furthermore, $E_f S(g) > E_f S(h) \Longleftrightarrow d_2(f,g) < d_2(f,h)$ for all $g, h \in B_2^+(f) \equiv \{g \in S \mid (f,g) \geq 0\}$. (See Figure 1 for illustration.)

Proof: $E_f S(g) \equiv (f,g) \leq \|f\|_2 \cdot \|g\|_2 = 1$, by the well known Schwarz inequality, with equality holding iff $f=g$, thus establishing the first part of the proposition.

For the second part, fix $f \in B_2$ and let $\pi : L_2 \to L_2$ be defined by $\pi g \equiv g - (f,g)f$ (orthogonal projection onto $\langle f \rangle^{\perp}$). Pick $g,h \in B_2^+(f)$ and note that $1 = \|g\|_2^2 = \|\pi g\|_2^2 + (f,g)\|f\|_2^2$, so $\|\pi g\|_2^2 = 1 - (f,g)$. Note also that $f-g = -\pi g + \left[1-(f,g)\right]f$, and therefore $\|f-g\|_2^2 = \|\pi g\|_2^2 + 1 - (f,g) = 2 - 2\,(f,g)$. Similarly, $\|f - h\|_2^2 = 2 - 2(f,h)$. The equivalences $E_f S(g) > E_f S(h) \Longleftrightarrow (f,g) > (f,h) \Longleftrightarrow ||f-g||_2 < ||f-h||_2 \Longleftrightarrow d_2(f,g) < d_2(f,h)$ are now immediate. QED.

Proposition 1 shows that the naive scoring rule would be effective (with respect to d_2) if $D \subset B_2$, which is unfortunately not the case. All is not lost, however; D is contained in $L_2 \setminus \{0\}$, which can be projected onto B_2 by the map ρ, where $\rho\, g \equiv g \,/\, \|g\|_2$ (See figure 1). This map induces a metric d on D, where $d(f,g) \equiv \|\rho f - \rho g\|_2$, as the following proposition shows:

Proposition 3: The function d above is a metric on D; i.e., for all $f, g, h \in D$, d satisfies the following three conditions:

(a) $d(f,g) \geq 0$; $d(f,g) = 0 \Longleftrightarrow f = g$.

(b) $d(f,g) = d(g,f)$ (symmetry)

(c) $d(f,g) \leq d(f,h) + d(g,h)$ (triangle inequality).

Proof: $d(f,g) \geq 0$ is immediate from the definition. Suppose $d(f,g) = 0$. Then by definition of 2-norm $\rho f = \rho g$ almost everywhere; i.e., $f(x) = c \cdot g(x)$ for almost every x, where the constant c is $\|f\|_2/\|g\|_2$. Since f and $g \in D$, we can integrate both sides and conclude $c = 1$ and therefore $f = g$. The symmetry of d is obvious, so there remains to establish only the triangle inquality $d(f,g) \leq d(f,h)+d(h,g)$. But this is an immediate consequence of the well-known Minkowsky inequality for the 2-norm. QED.

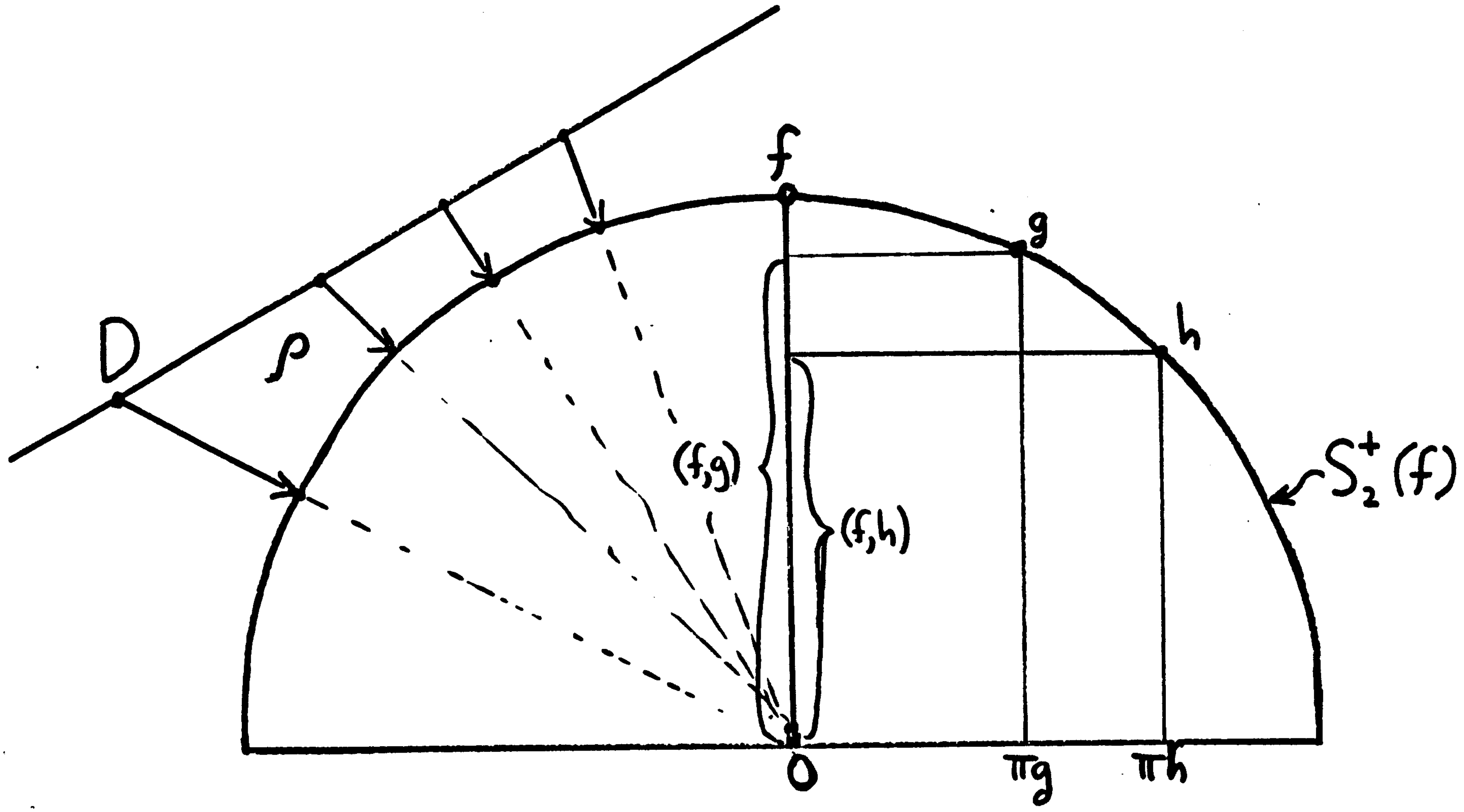
One can check with examples that the metric d accords well with most intuitive notions of similarity of density functions. This should come as no surprise, because d is the result of an appropriate "renormalization" of L_2 metric. Another advantage of d is that, unlike d_2, it is unaffected by a change of scale in Ω.

If the mapping ρ is applied to the naive scoring rule, the spherical scoring rule results. Since ρ always maps D into $B_2^+(f)$, we have the desired result:

Proposition 4. The spherical scoring rule $S(g,x) = g(x)/\|g\|_2$ is effective with respect to the metric d on all of D.

Propositions 1 and 4 may easily be generalized in several ways. The density functions in D can be defined on any measurable subset Ω of virtually any vector space. At the cost of some technical complication, D can represent probability distributions which do not necessarily have bounded and continuous density functions. The methods used above would seem valid as long as $L_2(\Omega)$ is dense in D. Therefore the spherical and quadratic scoring rules are effective even if the random variable to be forecast is discrete, or bounded, or vector-valued, etc.

Figure 1: Diagrams for Propositions 1&2.



IV. An Application

The senior management of a multinational corporation requires forecasts of foreign exchange rates from its currency experts. The forecasts are input to a large and varied set of decisions, some of which require more detail than "point" (i.e., expected value) forecasts. The format requested for the forecasts is that of a "histogram" (i.e., piecewise constant density function) with five or fewer steps (see Figure 2). The general form of the forecast is:

$$h(x) = \begin{cases} h_k \text{ if } x\varepsilon I_k, \; k=1,\dots,5 \\ 0 \text{ otherwise,} \end{cases}$$

with $\sum_{i=1}^{5} p_i = 1.00$, where $P_i = w_i h_i$ and w_i = width I_i, the intervals I_i being disjoint. The set D of the previous sections thus consists here of all densities of this form.

Under the naive scoring rule, the forecaster receives the score $S(h,x) = h(x) = h_k = p_k/w_k$ if the actual FX rate falls into the forecast interval I_k. Thus he is rewarded both for a tight forecast (narrow interval) and an accurate one (high probability interval). To the extent that the forecaster attempts to maximize his expected score, however, this rule will provide him incentive to overstate the probabilities of the perceived modal (most likely) outcomes and thus understate his perceived uncertainty, as a careful examination of example a of Section II will show. Such bias can be eliminated if management employs the spherical scoring rule

$$S(h,x) = h_k \Big/ \left(\sum_{i=1}^{5} h_i p_i\right)^{1/2} .$$

If, for some reason, the metric d_2 were thought to better represent closeness of approximation for densities, management could use the quadratic scoring rule $S(h,x) = 2h_k - \sum_{i=1}^{5} h_i p_i$.

Use of either of these effective scoring rules in most cases will not greatly change the relative scores from those arising from the naive rule, but as we have seen there will over the long run be differences in incentives. Only experience in using the various rules can determine if the naive rule induces sufficient bias[3] to justify the use of the more complicated effective rules.

As a numerical illustration, suppose forecaster A, a member of the "kitchen sink" school, after careful introspection produced the following forecast of the year-end pengo/US$ spot rate (graphed as h in Figure 2):

interval	probability
10-11 pengos/$	10%
11-13 "	30%
13-13.5 "	20%
13.5-15 "	30%
15-18 "	10% .

Forecaster B, an econometric model, after its keeper fed in assumptions about exogenous variables, produced the following alternative forecase (g in Figure 2):

interval	probability
9-11 pengos/US$	1%
11-13 "	15%
13-15 "	68%
15-17 "	15%
17-19 "	1%

[3] A simple way of detecting such bias is to look at the proportion of forecasts for which the actual FX rate ends up within the central 50% and 90% intervals of the histograms.

At the end of the year, the scorekeeper observes a spot rate of 12.21 pengos/$. Under the naive rule, A is assigned the score 15 (= .30 ÷ (13-11), times an arbitrary 100 to make for easier reading) and B the score 7.5. Since $||h||_2 \approx .45$ the $||g||_2 \approx .50$, the correction involved in the spherical scoring rule changes the relative scores only slightly: 33.7 for A, 14.8 for B. Likewise, the use of the quadratic rule (score 29.8 for A, 14.7 for B) also leaves the relative scores about the same in this case.

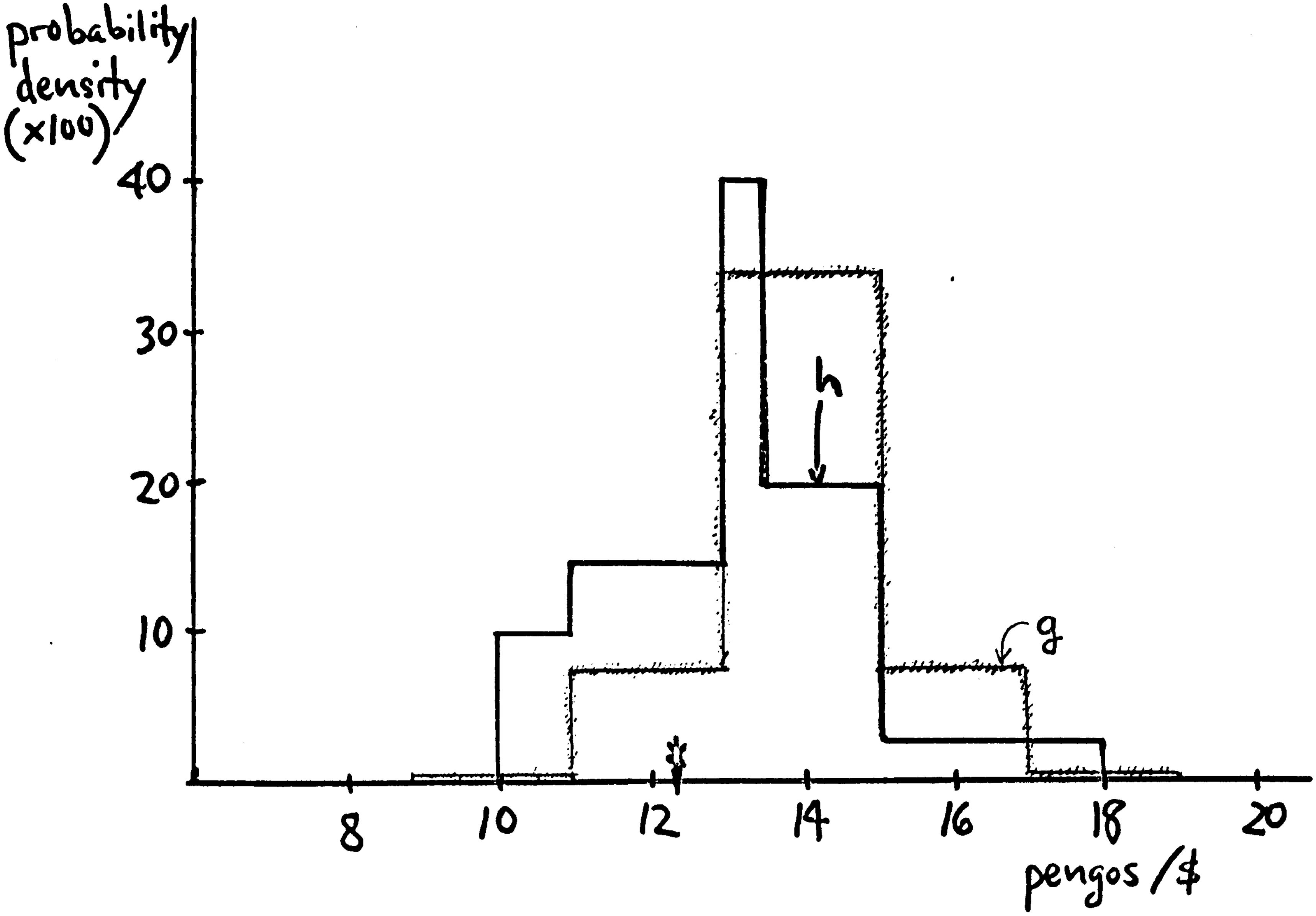
Little can be concluded on the basis of a single forecast. Although we couldn't observe the forecasters' "true" subjective probability distributions,[4] we can safely assume that they were well-approximated by the histogram forecasts, as long as the forecasters take their scores seriously. On the basis of a reasonably long "track record" of scored forecasts, management will be able to assess the merits of the two forecasters.

V. Discussion

There are other effective scoring rules besides those discussed here; Appendix A provides a characterization of those rules which are effective in some general sense. The virtue of the quadratic and spherical rules is not uniqueness, but simplicity: in a sense that Section III makes clear, these are the simplest scoring rules effective with respect to reasonable metrics (the metric d and the spherical rule being especially appealing in the author's opinion). In practical applications, simplicity is vital, and perhaps even more important than effectiveness, because a score provides little incentive to a forecaster who doesn't understand where it comes from.

[4] Note that even in the case of an econometric model forecaster, the forecast distribution depends on a subjective assessment of the distribution of the exogenous variables, and therefore can be regarded as subjective.

Figure 2: Histogram Forecasts



Another practical issue is that a forecaster may not actually attempt to maximize his expected score because he is risk averse. A practical solution to this problem is to use a rule such as the naive one, whose bias acts in a direction opposite to that of risk aversion by the forecaster. A theoretical resolution of this problem awaits further work.

So far we have cast the elicitation problem in a subjectivist mold: a scoring rule is supposed to give an expert incentive to introspect carefully, to make his beliefs explicit, and to accurately summarize them in a probability distribution F. There is also a more objectivist interpretation: Suppose there is essential uncertainty in the world, so that even a "perfect" expert (who employs accurate theory and has access to all current information and unlimited computing power) would provide a probability distribution F of positive variance for some event of interest. Suppose further that actual experts can approximate F to varying degrees, with better approximations generally costing more.

In this case, an effective scoring rule would motivate the expert to gather information and refine his approximate probability distribution G up to the point where his marginal expected gain due to improved score matches his marginal cost. If this marginal gain coincides with the marginal benefit to the decision makers, the rule is better than effective; it is, let us say, efficient. This notion of efficiency rejoins the communication problem to the decision problem from which it was separated in the introduction: efficiency is nothing more than effectiveness with respect to a metric based on the decision maker's loss function. It appears that slight generalizations of the spherical scoring rule will be efficient if the information gathering costs and the returns to informed decision making have sufficiently simple structure. The existence of efficient scoring rules in more general situations remains a very interesting open question.

Appendix A. Technical Notes

1. Generalized Versions of Known Scoring Rules.

The choice of scale for the score clearly does not affect the effectiveness of a scoring rule. If $S(g,x)$ is effective, then so is $aS(g,x) + b$ for any $a > 0$ and $b \in R$; this fact was used in the numerical example of Section IV. More generally, if ϕ is any strictly monotone increasing function, then $\phi \circ S$ is effective if S is, since effectiveness is an ordinal concept.

In many cases, one can use weighting functions to define generalizations of a given effective scoring rule. Recall that $\Omega \subset R$ is the set of possible outcomes (i.e., range) of the random variable, and let $W = \{w:\Omega \rightarrow R \mid w(x) \geq 0\}$ be some set of weighting functions. For any $w \in W$ and $g \in D$ (a set of density functions on Ω), define $w.g:\Omega \rightarrow R$ by $w.g(x) = \begin{cases} w(x)g(x)/\int_\Omega wf & \text{if } 0 < \int_\Omega wf < \infty, \\ 0 & \text{otherwise.} \end{cases}$

If -- as will often be the case -- $W.D \subset D$, then one can define the family $\{S_w\}_{w \in W}$ of scoring rules, where $S_w(g,x) = w(x)S(g,x)$. Assuming $S = S_1$ is effective, with respect to some metric d, S_w will be effective with respect to the metric d_w on $w.D \subset D$, where $d_w(f,g) = d(w.f, w.g)$. Such constructions may be useful in turning an effective scoring rule into an efficient scoring rule for certain types of loss functions. The idea of weighting scores is borrowed from Matheson-Winkler (1976).

2. Effectiveness in Matheson-Winkler.

Although their primary emphasis is propriety, Matheson and Winkler (1976) at two points touch on ideas related to effectiveness. Their discussion at the end of Section 2, in which they note a monotonicity property possessed by some of their scoring rules, has a spirit similar in our introductory remarks on effectiveness. Unfortunately, they define monotonicity relative only to a weak partial ordering, rather than a metric, and the ordering is outcome-dependent.

On the other hand, earlier in the same section, they show that a certain scoring rule is "effective" in the following sense: Let $C \subset L_2$ be a set of cumulative distribution functions (rather than density functions); call the rule $S:C \times \Omega \to R$ MW-effective if $E_F S(G) > E_F S(H) \Leftrightarrow d_2(F,G) < d_2(F,H)$, for all $F, G, H \in C$. Then their equation (24) shows that the rule $S(G,x) = -\int_{-\infty}^{x} G^2(t)dt - \int_{x}^{\infty}(1 - G(t))^2 dt$ is MW-effective. Although the MW concept of distance is weaker than ours (densities close implies CDF's close, but not conversely), it is quite compatible with our point of view, and provides a viable alternative definition of effectiveness.

3. A Characterization of Effectiveness.

We have usually discussed effectiveness with respect to some given metric. However, people may disagree over which metric best corresponds to intuitive ideas of distance between densities. Also, metrics that do not correspond to anyone's intuition may arise from considerations of the "efficiency" of scoring rules. Therefore, it makes sense to ask whether a scoring rule is effective with respect to any metric at all. Given some non-trivial set D of densities, call a rule G-effective ("generalized effective") if there is some metric d on D with respect to which it is effective.

Proposition 5 A scoring rule $S:D \times \Omega \to R$ is G-effective iff it is strictly proper and there is a function $\phi:D \times R \to R$ satisfying:

(a) for all $f \in D$, $\phi(f,\cdot) \equiv \phi_f$ is strictly monotone decreasing on $(m_f, M_f]$, where $m_f = \inf_{g \in D} E_f S(g)$ and $M_f = E_f S(f)$;

(b) For all $f \varepsilon D$, ϕ_f is non-negative and $\phi_f(M_f) = 0$;

(c) For all $f, g \varepsilon D$, $\phi_f(E_fS(g)) = \phi_g(E_gS(f))$; and

(d) For all $f, g, h \varepsilon D$,

$$\phi_f(E_fS(g)) \leq \phi_h(E_hS(f)) + \phi_h(E_hS(g)).$$

<u>proof</u>: First suppose S is G-effective. Then, for some metric d on D, and all $f, g, h \varepsilon D$,

(*) $E_fS(g) > E_fS(h) \Leftrightarrow d(f,g) < d(f,h)$.

Clearly, (*) implies that S is strictly proper. To derive ϕ, fix $f \varepsilon D$, and consider the functionals $A_f \equiv d(f,\cdot)$ and $B_f \equiv E_fS(\cdot)$ on D. Suppose $A_f(g) = A_f(h)$. Then (*) implies that $B_f(g) = B_f(h)$ and conversely; i.e., A_f and B_f have the same level sets in D. Therefore, there is some invertible real function ϕ_f: $\mathrm{im}B_f \to \mathrm{im}A_f$ such that $A_f = \phi_f \circ B_f$. From (*), we conclude that ϕ_f is strictly monotone decreasing on im $B_f \supset (m_f, M_f]$. Such a ϕ_f can be constructed for every $f \varepsilon D$ and (a) will be automatically satisfied. Since by construction $\phi_f(E_fS(g))$ is a metric, (b) - (d) will also be satisfied.

For the converse, assume S is strictly proper and ϕ satisfying (a) - (d) is given. Set $d(f,g) = \phi_f(E_fS(g))$. Then (b) - (d) imply that d is a metric, and (a) implies that S is effective with respect to d. QED

Proposition S can also be used to characterize those scoring rules effective with respect to a given metric d; one requires that the rule is proper and that there is some ϕ such that $d(f,g) = \phi_f(E_fS(g))$. For instance, if S is the quadratic scoring rule and $\phi_f(t) = ((f,f) - t)^{1/2}$, then $\phi_f(E_fS(g)) = d_2(f,g)$. Likewise, $\Psi_f(t) = (2 - 2t/\| f \|_2)^{1/2}$ links the spherical scoring rule to the "renormalized" L_2-metric.

4. <u>The logarithmic scoring rule</u>. Proposition S suggests that most proper scoring rules will not be effective with respect to any metric, no matter how

artificial, since a function ϕ with the required properties is usually hard to come by. Unfortunately, the characterization is not analytically very tractable. For instance, one would like to know if the logarithmic scoring rule $S(g,x) = \log g(x)$ is G-effective. To avoid the difficulties alluded to in Section II, let $\Omega = [0,1]$ and let D contain only densities on Ω bounded away from 0. I conjecture that there is no metric d on D with respect to which S is effective. Otherwise, by Proposition 5(c), for any $f,g \varepsilon D$, there would be strictly monotone functions ϕ_f, ϕ_g such that $\int f \log g = \phi_f^{-1} \phi_g(\int g \log f)$, which hardly seems possible. (The discrete version of this statement indicates for any p, q $\varepsilon \Sigma_n = \{(p_1,\ldots,p_n) | p_i > 0, \Sigma p_i = 1\}$, there is a way to transform $\prod_{i=1}^{n} p_i^{q_i}$ into $\prod_{i=1}^{n} q_i^{p_i}$ in two very simple steps, involving only the product, not the multiplicands.)

Ed Leamer points out that if this conjecture is true, it is a blow against maximum likelihood methods. The reasoning is as follows: Suppose D is a finite-dimensional set of densities, indexed by a vector θ of parameters, so $D = \{g(x;\theta) | \theta \varepsilon \Theta \subset R^n\}$. For example, D might be the histogram family of Section IV or the family of normal densities. Suppose one draws random samples from a distribution with some unknown density f. The log likelihood of the sample $\{x_1,\ldots,x_n\}$ is $L_N(\theta) = \sum_{i=1}^{N} \log g(x_i;\theta)$; as $N \to \infty$, $\frac{1}{N} L_N(\theta) \to I(\theta) = \int f(x) \log g(x;\theta)dx$. To estimate f by a density d ε D by the method of maximum likelihood, one picks the $g(x;\theta) \varepsilon D$ which maximizes $I(\theta)$. But $I(\theta) = E_f S(g(x;\theta))$, where S is the logarithmic scoring rule. If f ε D, the procedure is sound, since S is strictly proper. However, if -- as will often be the case -- $f \notin D$ and if the logarithmic scoring rule is not G-effective, then the g selected by the maximum likelihood procedure will not generally be the closest approximation to f in D in <u>any</u> (metric) sense of closeness.

5. Unresolved Mathematical Questions.

   a. Is the logarithmic scoring rule G-effective?

   b. Is there a more analytically useful characterization of G-effective scoring rules?

REFERENCES

Matheson, J. E. and Winkler, R. L., "Scoring Rules for Continuous Probability Distruibutions," Management Science, V. 22, No. 10, pg. 1087-1096, 1976.

Savage, L.J., "Elicitation of Personal Probabilities and Expectations," Journal of the American Statistical Association, V. 66, No. 366, December 1971.

Rudin, W., Real and Complex Analysis, McGraw Hill, N. Y. 1966.