

Pre-Verification Scoring Rules for the Thoughtful Evaluation of Ideas

Jeffrey E. Danes

**Orfalea College of Business, Marketing
California Polytechnic State University
San Luis Obispo**

Abstract

Traditional scoring rules provide incentives for eliciting subjective probability, degrees of personal belief, and payoffs for probabilities that accurately predict the outcome of some event, such as the success of an innovative but untried idea. Newly developed weighted scoring rules by Jose, Nau, and Winkler (2008) are used to develop pre-verification rules to deliver payoffs when the outcome of the event has not yet occurred. Personal probabilities are assumed to be beta distributed as is the consensus distribution, the aggregation of the individual's probabilities. Two methods of aggregating probabilities into a consensus distribution are discussed: Winkler's (1968) Bernoulli model and a new model which we call the centered Bernoulli model. The Condorcet Jury Theorem and centered model are used to create a target distribution that corresponds to the probability that untried idea will or will not be successful. A person whose personal probability is close to the target receives a larger payoff.

Key words: subjective probability, scoring rules, weighed scoring rules, utility, Condorcet Jury Theorem, incentive compatibility, aggregation of probabilities, consensus distribution, wisdom of crowds

1. Introduction

Perhaps the best way to reward customers or employees for their accuracy in evaluating the probable success of a new idea is to try the idea and see if it performs as anticipated. However, what if success is only realizable in say 5 or 10 years? Or, what if empirical evaluation of all suggestions is prohibitively expensive? This paper develops a pre-verification scoring rule that provides near immediate payoffs to customers or employees for their thoughtful evaluations of an idea when the idea itself has not yet been empirically evaluated. The thoughtful evaluation of ideas is relevant to managers who collect suggestions from their employees, marketers who collect new product concepts from their customers and community planners who collect proposals from their citizens. The purpose of this paper is to develop a scoring rule to reward a person (expert or assessor) for the thoughtful evaluation of an idea without requiring the empirical verification of the event.

In an attempt to keep the meaning of an idea relatively clear, on one hand, our reference to an idea is analogous to the concept of a new product as defined by Crawford and Benedetto (2006). A product concept includes technology, form, and benefit. If the idea is physical then the form is physical, if the idea is an intangible service, then the form is a series of steps designed to deliver the service. In both cases, physical or service, the technology enables the form to deliver a benefit. On the other hand, an idea could be some course of action, a proposed solution, or an event. By an evaluation of an idea, we mean the probability of success, the probability that the idea will be successful on some unambiguous, relevant dimension or set of dimensions. The following two questions illustrate the intent: “In your opinion, what is the

probability that product X will be the market share leader within the next 3 years? “Or “In your opinion, what is the probability that both polar caps will melt by the year 2050? “¹

The evaluation measure for the pre-verification scoring rule is the probability forecast of how good an idea is relative to some acceptable measure of success. The paper is organized as follows. Section 2 briefly reviews traditional and weighted scoring rules and introduces the problem. Section 3 discusses a Bayesian model proposed by Winkler (1968) and develops a revised, centered, model for aggregating individual probabilities into a consensus distribution. Section 4 defines expertise as a characteristic of probability and discusses two versions of the Condorcet Jury Theorem. Section 5 presents the model for the pre-verification scoring rule. One is given without weights of evidence and one is given with weights of evidence. Section 6 concludes with a summary overview of the paper’s contributions. We assume an individual’s perception of the quality of an idea has some degree of validity, the quality of the idea does not vary over the time of assessment, and that individuals desire to maximize utility. Other assumptions are introduced as needed.

2. Scoring Rules

Traditional Scoring Rules

We first discuss traditional scoring rules followed by weighed scoring rules. Traditional scoring rules provide incentives and payoffs for eliciting subjective probability, degrees of personal belief (de Finetti 1937; Brier 1950; Good 1952; Winkler and Murphy 1968; Savage 1971; Matheson and Winkler 1976; Winkler 1996; Selten 1998). Let r_i denote a person’s revealed or reported subjective probability, p_i be a person’s true probability, and e denote a binary event

¹ Probability can be assessed in number of ways; the approach featured in this paper is the ratio of success to successes and failures, $s/(s+u)$.

with indicator 1 if the event occurs and 0 if the event does not occur. Scoring rules may be used *Ex Ante*; that is, to encourage a person to give an honest probability judgment, provide an honest report of their true probability. *Ex Ante* scoring rules are designed so that if the person's goal is to maximize expected utility, then their best strategy is to report, $r_i = p_i \sim e$ where p_i is their best estimate of e . A scoring rule used *Ex Post* rewards a person for the accuracy of a probability judgment (Winkler 1996).

Define $S(r)$ to be a scoring rule that gives the person a payoff $S_1(r)$ if the event occurs and payoff $S_2(r)$ if the event does not occur. The person's expected payoff for this binary situation is $E[S(r)] = p S_1(r) + (1-p) S_2(r)$, where p is the person's true probability. The scoring rule is *strictly proper* if $E[S(p)] > E[S(r)]$ for $r \neq p$. In the *Ex Ante* case, the assessor anticipates the payoff and thus there is an incentive to provide the best possible probability judgment, $r = p$. In the *Ex Post* case, the assessor is rewarded for accurate for probability judgments. Traditional scoring rules assume a payoff follows only after the event has occurred. However, how does one allocate rewards *before* the event occurs? Winkler (1999) has suggested the use of the expected payoff, $E[S(r)] = \bar{p} S_1(r) + (1-\bar{p}) S_2(r)$, where \bar{p} is now some other useful, available information used to estimate e . The challenge here is in obtaining the other useful information. The proposal in the present paper is to model an intermediate step q that intervenes between p_i and e . That is, we seek a pre-verification probability density function, *pdf*, that precedes the event whose expectation, q , is a better estimate of the event e than is p_i .

$$r_i = p_i \sim q \sim e \quad (1)$$

Scoring rules also apply to discrete and continuous distributions, the expected scoring rule for discrete distribution is

$$E[S(r_1, r_2, \dots)] = \sum_{j \in I} p_j S_j(r_1, r_2, \dots).$$

And the expected score is strictly proper if

$$E[S(p_1, p_2, \dots)] > E[S(r_1, r_2, \dots)] \text{ , where } r_k \neq p_k \text{ for any } k \in I.$$

The three standard forms for discrete distributions are provided in Winkler (1996)

$$\text{Quadratic: } S_j(r_1, r_2, \dots) = 2r_j - \sum_{k \in I} r_k^2$$

$$\text{Spherical: } S_j(r_1, r_2, \dots) = r_j / (\sum_{k \in I} r_k^2)^{1/2} \quad (2)$$

$$\text{Logarithmic: } S_j(r_1, r_2, \dots) = \text{Log } r_j$$

The three standard forms for continuous distributions are given in Matheson and Winkler (1976).

The discrete rules may be extended to the continuous case by limiting arguments.

$$\text{Quadratic: } S(r(x)) = 2r(x) - \int_{-\infty}^{\infty} r^2(x) dx,$$

$$\text{Spherical: } S(r(x)) = r(x) / (\int_{-\infty}^{\infty} r(x)^2)^{1/2} \quad (3)$$

$$\text{Logarithmic: } S(r(x)) = \text{Log } r(x)$$

Weighted Scoring Rules

Jose, Nau, and Winkler (2008) in working with weighted forms of pseudospherical and power scoring rules show that an assessor's expected score is a minimum when p is a uniform distribution. Thus expected scores reward the assessor in proportion to the divergence (some

measure of “distance”) of p from a uniform distribution. They argue that in most applications the relevant reference point is not a uniform distribution; accordingly, they generalize scoring rules to reward the forecaster in proportion to the divergence of p from an appropriate baseline distribution q , the weighted scoring rule. The expected weighted score is denoted $S(r, p|q)$ and the optimal weighted expected score is $S(p, p|q)$ or $S(p|q)$ for simplicity. In their work, both the quadratic and spherical rules are normalized by an affine transformation, $(p_i/q_i)^{\delta-1}$ and generalized from the Euclidian 2-norm to the δ -norm, which for the spherical rule, gives the pseudospherical rule proposed by Good (1971). In their analytic work, Jose, Nau, and Winkler (2008) discovered that the weighted power expected score is identical to directed (power) divergence proposed by Havarda and Chavrat (1967), the weighted pseudospherical expected score is the cross entropy measure introduced by Arimoto (1971) and that both of these generalized divergences reduce to the well known Kullback-Liebler divergence (Kullback 1959) when δ -norm = 1. The Kullback-Liebler divergence corresponds to a weighted log scoring rule. The Kullback-Liebler divergence for discrete probabilities is defined as

$$S(p|q) = \sum_k (p_k \text{Log } p_k - p_k \text{Log } q_k) = \sum_k (p_k \text{Log } \frac{p_k}{q_k}) \quad (4)$$

In general $S(p|q) \geq 0$ and does not necessarily satisfy the triangle inequality axiom and hence is not a true distance metric, $S(p|q) \neq S(q|p)$ (Cover and Thomas 1991). Of the various weighted scoring rules investigated by Jose, Nau, and Winkler (2008, p. 00), they claim, “A well-chosen and not-necessarily uniform baseline distribution is the most important parameter of the scoring rule in any case.” The traditional logarithmic scoring rule reported above in (2) and (3) above has been criticized by Selten (1998) on the grounds that it is hypersensitive to small probabilities

and is not enough sensitive in discriminating between true and reported probabilities. This deficiency does not exist with the weighted logarithmic scoring rule. The right hand side of equation (4) reveals an important ratio, (p_k/q_k) ; thus the weighted logarithmic rule is influenced by relative rather than absolute probabilities. Other benefits of scoring rules based upon the ratio (p_k/q_k) are discussed in Jose, Nau, and Winkler (2008).

The forecast distribution or *pdf* and the target distribution or *pdf* for a weighted expected scoring rule is a relevant pair of discrete or continuous distributions. Let $p(x)$ and $q(x)$ be some continuous distributions, then the Kullback-Liebler divergence is defined as

$$S(p \parallel q) = \int_{\mathcal{X}} p(x) \text{Log } p(x) - p(x) \text{Log } q(x) d\mu(x) \quad (5)$$

Where the distribution on the outcome space \mathcal{X} is absolutely continuous with respect to a σ -finite probability measure, μ .

3. Aggregating Probability Distributions

As noted above, our goal is two-fold, first we seek a probability, q , which is better estimate of e than is p_i and second we seek a scoring rule for $r_i = p_i \sim q \sim e$. Toward this end, we discuss the work by Winkler (1968) who provides a Bayesian model for the aggregation of probabilities into a consensus distribution. We then discuss a new way to estimate the weights in Winkler's (1968) model; this newer model is called, the *centered* Bernoulli model.

Consider a sequence of 'successes and 'failures' to reflect the parameters of an imaginary binomial distribution, where $x = [0,1]$ with probability, $0 < p < 1$. The observation of a single random variable, \tilde{x} , is a 'trial' or a 'point'. This model capitalizes on de Finetti's Theorem that

shows that one can treat the sequence of 1s and 0s *as if* their densities were averaged over p . The probability p is treated as a random variable and the natural conjugate of a Bernoulli process is the beta distribution. Define p_i as the expected probability for person i ,

$$p_i = \frac{s''}{s''+u''} ,$$

where, s'' is the number of success points, u'' is the number of failure points, and $\varphi = s'' + u''$ is person's total points. Let the following prior probability for person i be random a variable and be beta distributed,

$$f(p_i | s, u) = \frac{\Gamma(s+u)}{\Gamma(s)\Gamma(u)} p_i^{s-1} (1-p_i)^{u-1} . \quad (6)$$

The posterior distribution is given following Raffia and Schlaifer (1961).

$$f(p_i | s'', u'') = \frac{\Gamma(s''+u'')}{\Gamma(s'')\Gamma(u'')} p_i^{s''-1} (1-p_i)^{u''-1} \quad (7)$$

where $s'' = s' + s$ and $u'' = u' + u$ and s' represents the data, the intervening number of successes.

The expected probability is the probability assessor's true probability, p_i

$$E[f(p_i | s'', u'')] = \frac{s''}{s''+u''} = p_i \quad (8)$$

The Bernoulli Model

Winkler (1968) developed one of the first Bayesian models for the aggregation of probabilities.

Let us define multiple sources of information as s''_i to represent the input from a group of n people (or experts) to a decision-maker as follows,

$$S = \sum_i \omega_i s''_i + 1 , \quad (9)$$

$$U = \sum_i \omega_i \mu''_i + 1,$$

Winkler's (1968) model for a consensus probability distribution is given as the following posterior *pdf* with S and U serving as the beta parameters.

$$f(p|S,U) = \frac{\Gamma(S+U)}{\Gamma(S)\Gamma(U)} p^{S-1} (1-p)^{U-1} \quad (10)$$

Winkler (1968) notes that the decision-maker needs to decide on how to weight the input from others. If the information is completely redundant, then $\sum_i \omega_i = 1$ and if the information is independent, then $\sum_i \omega_i = n$ people. Winkler (1968) treats the weights as a heuristic device to demonstrate the influence that information dependence has on the variance of the beta distribution. The restrictions stipulated by Winkler (1968) will be relaxed when the centered model is presented. The expected probability is a weighted group mean,

$$E_g[f(p|S,U)] = \frac{S}{S+U} = \bar{p}_w, \quad (11)$$

where E_g stands for expectation over the group.

A Centered Bernoulli Model

The centered Bernoulli model assumes the mean of the beta distribution is \bar{p} rather than \bar{p}_w .

Empirical research in forecasting has found simple averages to predict remarkably well (Clemen and Winkler 1999; 2007). Clemen (1986) for example in reviewing the forecasting literature on pooling point forecasts somewhat disappointingly reports that simple aggregation methods work as well as the theoretically derived, complex methods, and sometimes better. Armstrong (2001) in a re-analysis of 30 studies finds the simple average to improve forecast accuracy from 3.4% to 23.5% relative to the mean performance of the forecasts being averaged. Clemen and Winkler

(1999, 2007) report essentially the same when reviewing various methods of aggregating probability distributions. The simple average forecast probability is very robust. When used as a forecast probability, empirical research tends to support \bar{p} over other more complex models. We make use of this empirical finding below.

Similar to Winkler's (1968) model we assume individual probabilities p_i are beta distributed,

$$E[f(p_i | s'', u'')] = \frac{s''}{s'' + u''} = p_i \quad (12)$$

The primary task for the centered Bernoulli model is to maximize the comparability of the density function for p_i and the density function for p . In other words, if $\bar{p} = p_i = p_l$, then we desire $f(\bar{p}) = f(p_i) = f(p_l)$. We show below that a centered Bernoulli model accomplishes this objective.

$$E_g[f(p | a, b)] = \frac{a}{a + b} = \bar{p} \quad (13)$$

Our desire is to use \bar{p} rather than \bar{p}_w as the mean of the beta distributed probabilities. Toward this end, let the beta parameters (a, b) be those estimated from group data as follows,

$$a = \bar{p} \left[\left(\frac{\bar{p}(1 - \bar{p})}{\text{var}(p)} \right) - 1 \right] \quad \text{and} \quad b = (1 - \bar{p}) \left[\left(\frac{\bar{p}(1 - \bar{p})}{\text{var}(p)} \right) - 1 \right]. \quad (14)$$

The goal is to make the densities $f(p_i)$ and $f(p)$ comparable and do so we define weights ω_i so

that $E_g(\omega_i s'') = a$ and $E_g(\omega_i u'') = b$. One specification of ω that does exactly this is given as

α and β below,

$$\alpha = \left(\frac{a+b}{s''+u''} \right) s'', \quad \alpha = \omega s'',$$

$$\beta = \left(\frac{a+b}{s''+u''} \right) u'', \quad \beta = \omega u'' \quad (15)$$

The individual's parameters $(\omega s'', \omega u'')$ are denoted (α, β) with weights defined as follows, $\omega = (a+b)/(s''+u'')$. The goal is to transform the individual's parameters (s'', u'') to a form that is comparable to the center, the group (a, b) . Comparability in this context means,

$$E(\alpha) = a \text{ and } E(\beta) = b \text{ (proof omitted)} \quad (16)$$

It can also be shown that the transformations $(\omega s'', \omega u'')$ yield an equality of sums and means

$$(a+b) = (\alpha + \beta) \text{ and } a/(a+b) = \sum \alpha / (\sum \alpha + \sum \beta). \quad (17)$$

Thus, the mean probability obtained directly from the group is equal to mean of the posterior distribution obtained by updating the center's posterior via Bayesian updating. The resulting model for person i is

$$f(p_i | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1} (1-p_i)^{\beta-1} \quad (18)$$

The model defined in (18) has the following desirable properties.

$$E_g [f(p | E_g(\alpha), E_g(\beta))] = E_g [f(p | a, b)] = E_g [f(p)] = \bar{p} \quad (19)$$

The only constraint on the weights is that their sum be greater than 0, $\sum_i \omega_i > 0$. Note that this constraint is less restrictive than the constraint suggested by Winkler (1968). Table 1 below presents an example data set for 10 hypothetical people.

Table 1: Example Data

i	s''	u''	φ	p	ω	α	β	$\alpha - a$	$\beta - b$	$S(p \parallel \hat{q})$
1	1	1	2	0.50	2.85	2.85	2.85	-1.18	1.18	3.86
2	40	70	110	0.36	0.05	2.08	3.63	-1.96	1.96	2.39
3	8	5	13	0.62	0.43	3.51	2.20	-0.52	0.52	5.50
4	40	9	49	0.82	0.11	4.66	1.05	0.62	-0.62	10.77
5	36	2	38	0.95	0.15	5.41	0.30	1.37	-1.37	26.15
6	65	10	75	0.87	0.07	4.95	0.76	0.91	-0.91	13.59
7	100	50	150	0.67	0.03	3.81	1.90	-0.23	0.23	6.43
8	90	35	125	0.72	0.04	4.11	1.60	0.07	-0.07	7.61
9	20	2	22	0.91	0.25	5.19	0.52	1.15	-1.15	17.75
10	40	20	60	0.67	0.09	3.81	1.90	-0.23	0.23	6.43
Mean				0.707		$a=4.038$	$b=1.672$	0.00	0.00	
Sum	S=440	U=204			4.13					

Using Table 1, one can clearly see the similarities and differences between the Bernoulli model and the centered Bernoulli model. For Winkler's model the group parameters are (S, U) and the individual's parameters are (s'', u'') . For the centered Bernoulli model, the group parameters are (a, b) and the corresponding individual's parameters are (α, β) . For Winkler's Bernoulli model independence is assumed, that is, each weight is set to 1 so that $\sum_i \omega_i = 10$. In the centered

Bernoulli model the $\sum_i \omega_i = 4.13$, this number is shown in Table 1 under ω . The mean probability for Winkler's Bernoulli model is .683 and the mean probability for the centered model is .707. From the method of moments, using equations (14) $a = 4.038$ and $b = 1.672$. The individual's centered beta parameters as given by (18) are listed in Table 1 under α , and β . The corresponding density plots for the 10 individuals are presented below in Figure 1.

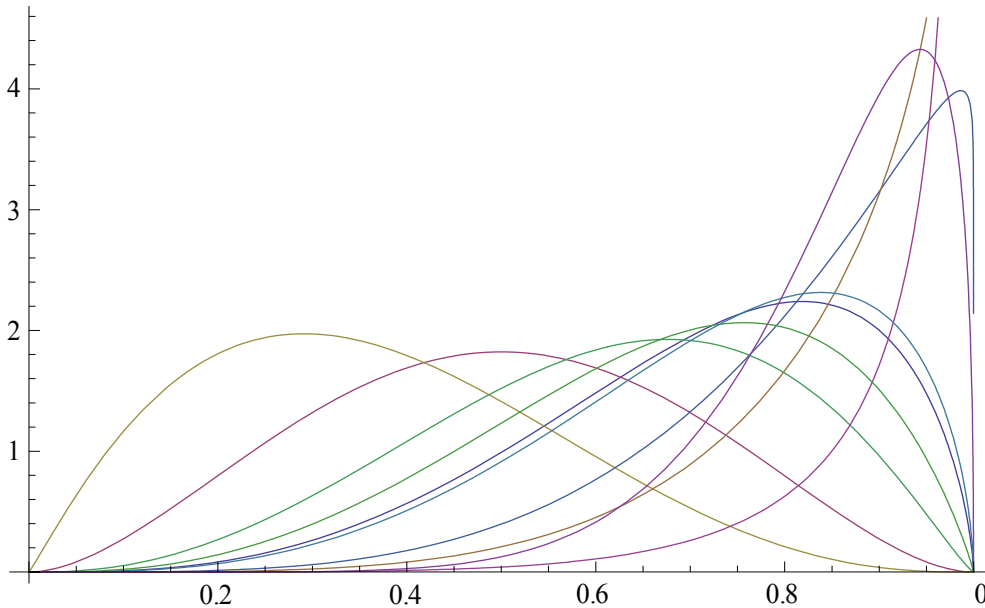


Figure 1. Density plots for the centered model, see Table 1

4. Probability, Expertise and Condorcet Jury Theorems

Expertise and Weight of Evidence

The probabilities in Table 1 are the ratios of successes to successes plus failures, $s''/(s''+u'')$. If a person reports more success and fewer failures, they are in a sense, revealing a type of awareness or knowledge. Hence, we can view probabilities that depart from 0.50 as expressing some type of knowledge or expertise. In probability forecasting, forecasts near 1 (or 0) are

called refined or *sharp* (Winkler 1986, DeGroot and Fineberg 1992). That is, if $|p_i - 0.5| > |p_l - 0.5|$ we can then say that person i is sharper than person l .

Another type of ‘knowledge’ is the weight of the evidence, φ , used in forming the probability, $\varphi = s'' + u''$, see Table 1. If $\varphi_i > \varphi_l$, we can then say that person i has more ‘evidence’ than person l . Weight of evidence will be addressed in Section 5.

Condorcet Jury Theorems

The key idea is that p_i may be interpreted as some degree of expertise. If the expertise of the members of group is greater than 0.50, $p > 0.50$, then when given the choice between two alternatives, the probability that a majority will select the correct option increases as group size increases. In the limit, the majority always makes the correct decision. Below two relevant Condorcet Jury Theorems are presented, see Grofman, Owen, and Feld (1983) for a review of 13 theorems.

Condorcet Jury Theorem, for Constant $p > 0.50$. Let $(\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_n)$ be an exchangeable sequence of binary independent random variables and let $p(z_i = 1) = p_i = p > 0.5$ for all i , and let $p_n = p(\sum_i \tilde{z}_i > n/2)$. Then $p_n > p$ and $p_n \rightarrow 1$ as $n \rightarrow \infty$.

Convergence to the asymptote is quick. For example if $p = 0.80$, then $p_{13} = 0.99$. The assumptions underlying the above theorem are: 1. there is a correct alternative; 2. the individual judgments are independent, stem from independent information, and 3. $p > 0.50$. Ladha (1992) discusses (1) the first assumption, there is a correct alternative. In brief, Ladha (1992) argues for

two cases where this assumption applies, state of the world, and specific types of social choice. In state of the world, a demonstrative example is whether a well will or will not produce water. In social choice, a relevant occasion is when given full information the choice is clear, say the target market will adopt the innovation, however, the group may be operating with limited information. (2) Recent work by Ladha (1992) and Berg (1993) relax the independence assumption. Both independently prove a version of the CJT when shared information is positively correlated. Positive correlation tends to impede convergence. Berg (1993) who also studied the effects of negatively correlated information, surprisingly, found that a small amount of negative correlation facilitates convergence. (3) On the last assumption of the three, Grofman, Owen, and Feld (1983) provide a generalized CJT for the average probability, \bar{p} , and this is provided below.

Condorcet Jury Theorem, for $\bar{p} > 0.50$. If $\bar{p} \leq 0.50$ then as $n \rightarrow \infty$, $p_n \rightarrow 0$;

If $\bar{p} > 0.50$ then as $n \rightarrow \infty$, $p_n \rightarrow 1$; while if $\bar{p} = 0.50$, $1 - \epsilon^{1/2} < p_n < \epsilon^{1/2}$, that is,

$$0.39 < p_n < 0.61.$$

This result shows that if the mean probability \bar{p} is > 0.50 the theorem then applies to any distribution, regardless of the shape of the distribution. CJT has also been extended to categorical and interval scales.

5. Pre-Verification Scoring Rules

The Condorcet jury theorem, under the conditions specified, states that probability of the group selecting the correct alternative approaches one as the size of the group increases. Thus, an important implication of the generalized CJT is that for a group with $\bar{p} > 0.50$, when n is

sufficiently large the event, $e = 1$. This notion for a finite group is captured in equation (20) for events that have not yet occurred.

$$\hat{e} \approx \begin{cases} 1 & \text{if } \bar{p} > 0.50 \\ 0 & \text{if } \bar{p} < 0.50 \end{cases} \text{ and } n < \infty, \quad (20)$$

where \hat{e} is used instead of e to denote an unobserved event. The symbol \approx is used to denote that since the actual event has not yet occurred, there will always be some uncertainty. We are now in a position to define the appropriate q for $r_i = p_i \sim q \sim e$, where q is an estimate of e . For this purpose we use the individual's centered beta distributions in equation (18). In particular, we seek the a scoring rule for the following,

$$f(r_i) = f(p_i) \sim f(q), \quad (21)$$

where the expectation $E_g f(q) = \hat{e}$, and $0 < \hat{e} < 1$.

The intuition behind the pre-verification scoring rule is seen by inspecting the following beta distributions obtained from the centered Bernoulli model in equation (18) with arbitrary parameters.

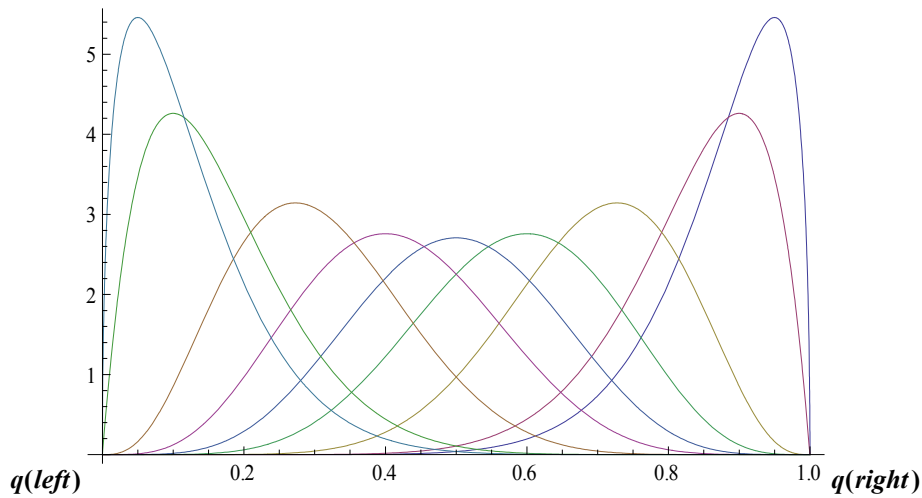


Figure 2. Various Target Distributions

According to (20) above, the specification of $f(q)$ in (21) depends upon the value of \bar{p} , as follows:

$$f(r_i) = f(p_i) \sim f(q(right)), \text{ if } \bar{p} > 0.50$$

$$f(r_i) = f(p_i) \sim f(q(left)), \text{ if } \bar{p} < 0.50 \quad (22)$$

Reasonable values for $f(q(right))$ are to select $\hat{\alpha} >$ all other α and $\hat{\beta} <$ all other β and corresponding values for $f(q(left))$ are to select $\hat{\alpha} <$ all other α and $\hat{\beta} >$ all other β . This rule is applied with the constraint that sum of any pair of beta parameters be equal to the sum of any other pair of beta parameters, be it group, individual, or modified: $(a+b) = (\alpha_i + \beta_i)$

$= (\alpha_i + \beta_i) = (\hat{\alpha} + \hat{\beta})$. Another reasonable constraint is to restrict the probability $\hat{\alpha} / (\hat{\alpha} + \hat{\beta})$ to be .95 or greater for $f(q(right))$ or .05 or less for $f(q(left))$.

The parametric version of the Kullback-Leibler divergence (Kullback, 1959) for beta distributed probabilities is:

$$S(p \parallel \hat{q}) = \ln \left\{ \frac{B(\alpha, \beta)}{B(\hat{\alpha}, \hat{\beta})} \right\} - (\alpha - \hat{\alpha}) \psi(\alpha) - (\beta - \hat{\beta}) \psi(\hat{\beta}) + \psi(\hat{\beta}) + (\alpha - \hat{\alpha} + \beta - \hat{\beta}) \psi(\alpha + \beta), \quad (23)$$

where $\psi(\cdot)$ is the digamma function and $\hat{\alpha}$ and $\hat{\beta}$ are defined below. Above it was noted that the Kullback-Liebler divergence, $S(p|q)$ is ≥ 0 . Thus an application of scoring rules based upon (22) above should consider whether smaller a score is the better estimate of \hat{e} or whether a larger score is the better estimate of \hat{e} . The discussion below assumes a larger score is better.

As such the set up for the scoring rule is as follows. If $\bar{p} > .50$ then $f(q(right))$ is an appropriate

target distribution with scoring rule $S(p \parallel \hat{q}(left))$. And, if $\bar{p} < .50$ then $f(q(left))$ is an appropriate target distribution with scoring rule $S(p \parallel \hat{q}(right))$. This setup insures that greater divergence reflects greater congruence with the target distribution and a $S(p \parallel q) = 0$ indicates the worst possible score.

The example provided in Table 1 has sum of parameters, $(a + b) = (\alpha_i + \beta_i) = 5.71$ with $\bar{p} = .707$. Thus we desire the rule $S(p \parallel \hat{q}(left))$ so that greater divergence means an individual's density $f(p_i)$ is divergent from $f(q(left))$, closer to the target density $f(q(right))$. The right most column in Table 1 reports the Kullback-Liebler divergence, $S(p \parallel \hat{q}) = S(p \parallel \hat{q}(left))$, as given by equation (23). In this example, we applied a 5% rule so that $\hat{\alpha} = .286$ and $\hat{\beta} = 5.425$; note that $\hat{\alpha}$ is smaller than all other α and that $\hat{\beta}$ is larger than all other β , as suggested above, see Table 1.

Pre-Verification Scoring Rules with Weights of Evidence

Thus far have weight of evidence, $\varphi = s'' + u''$, have been ignored and they are easily incorporated into the pre-verification scoring rule. $\varphi = s'' + u''$ reflects the number 'trials' or 'points' and if $\varphi_i > \varphi_l$ we can then say that person i has more 'evidence' than person l . A reasonable way to employ weight of evidence in the scoring rule is to amplify the payoff in proportion to the evidence.

$$Payoff_i = K (\varphi_i S(p \parallel \hat{q})), \quad (24)$$

where K is an arbitrary constant of proportionality. A linear transformation of a scoring rule is itself a scoring rule so equation (24) remains a scoring rule.

6. Summary

This paper develops a pre-verification scoring rule for the evaluation of ideas; the evaluation measure is a probability forecast of how good an idea is relative to some acceptable measure of success. Traditional scoring rules provide incentives for eliciting probabilities and payoffs for probabilities that accurately predict the outcome of some event, such as the success of a new product concept. In the *Ex Ante* case, the assessor anticipates the payoff and thus there is an incentive to provide the best possible probability judgment, the reported probability is the person's true probability, $r_i = p_i$. In the *Ex Post* case, the person is rewarded for accurate for probability judgments. Traditional scoring rules assume payoffs follow only after the event has occurred, for example, after the new product is deemed successful.

The proposal in the present paper is to model an intermediate step q that intervenes between p_i and the event, e . That is, we seek a pre-verification probability density function, *pdf*, that precedes the event whose expectation, q , is a better estimate of the event e than is p_i , $r_i = p_i \sim q \sim e$. The new scoring rule provides near immediate payoffs to assessors for their thoughtful evaluations of ideas when the ideas have not yet been empirically evaluated.

Two methods of aggregating probabilities into a consensus distribution are discussed: Winkler's (1968) Bernoulli model and a new model which we call the centered Bernoulli model. The Condorcet Jury Theorem and centered model are used to create a target distribution that corresponds to the probability that untried idea will or will not be successful. A person whose personal probability is p_i close to the target distribution receives a larger payoff.

References

- Arimoto, Suguru (1971), Information-Theoretical Considerations on Estimation Problems, *Information and Control*, 19(3), 181-194.
- Armstrong, J. Scott (2001), Combining Forecasts, J.S. Armstrong, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer, New York.
- Bickel, J. Eric (2007), Some comparisons among quadratic, spherical, and logarithmic scoring rules, *Decision Analysis*, 4(2), 49-65.
- Berg, Joyce E., Robert Forsythe, Forest D. Nelson, and Thomas A. Rietz (2001), Results from a dozen years of election futures markets research, still forthcoming in Charles A. Plott and Vernon Smith, eds., *Handbook of Experimental Economic Results*, Elsevier, Amsterdam.
- Chen, Yiling, Chao-Hsien Chu, Tracy Mullen, and David M. Pennock (2005), "Information markets vs. opinion pools: An empirical comparison," *Proceedings of the 6th ACM conference on Electronic commerce*, Vancouver, British Columbia, Canada.
- Clemen, Robert T. (1989), Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting*, 5, 559-583.
- Clemen, Robert T. (2002), Incentive contracts and strictly proper scoring rules, *Test*, 11(1), 167-189.
- Clemen, Robert T. and Robert L. Winkler, (1990), Unanimity and compromise among probability forecasters, *Management Science*, 36, 767-779.
- Clemen, Robert T. and Robert L. Winkler, (1999), Combining probability distributions from experts in risk analysis, *Risk Analysis*, 19, 187-203.
- Clemen, Robert T. and Robert L. Winkler, (2007), Aggregating probability distributions, Ward Edwards, Ralph F. Miles Jr. Detlof, and von Winterfeldt, eds., *Advances in Decision Analysis From Foundations to Applications*, 154-176, Cambridge University Press.
- Condorcet, Marquis de. (1976), *Condorcet: Selected Writings*, Keith Michael Baker, ed., Indianapolis: Bobbs-Merrill. (Originally published in 1785).
- Cover, Thomas M. and Joy A. Thomas (1991), *Elements of Information Theory*, New York, Wiley-Interscience.
- de Finetti, Bruno (1937), La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7, 1-68. (Translated in 1980 by H. E. Kyburg, Jr., Foresight. Its logical laws, its subjective sources. H. E. Kyburg, Jr. & H. E. Smokler, eds., *Studies in Subjective Probability*, 2nd ed., 53-118. Huntington, New York: Robert E. Krieger.)

- de Finetti, Bruno (1974), *Theory of probability*, Vol. 1. Wiley, New York.
- de Finetti, Bruno (1977). Probabilities of probabilities: a real problem or a misunderstanding? in A. Aykac and C. Brumat, (Eds.) *New Developments in the Application of Bayesian Methods*, Amsterdam: North Holland Publishing Company.
- DeRosa, Darleen M., Carter L. Smith, and Donald A. Hantula, The medium matters: Mining the long-promised merit of group interaction in creative idea generation tasks in a meta-analysis of the electronic group brainstorming literature, *Computers in Human Behavior*, 23 (2007) 1549–1581
- Genest, Christian and James V. Zidek (1986), Combining probability distributions: A critique and annotated bibliography. *Statistical Science*, 1, 114-148.
- Geneiting, Tilmann and Adrian E. Raftery (2007), Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, 102, 359-378.
- Good, Irving J. (1985), Weight of evidence: A brief survey, *Bayesian Statistics 2*, pp 249-269, J.M. Bernardo, M.H. DeGroot, D.V Lindley, A.F.M. Smith, eds., North Holland, Elsevier Science.
- Good, Irving J. (1971), Comment on paper by Buehler, 337-339, V.P. Godambe and A. Sprot, eds., *Foundations of Statistical Inference*, Holt, Reinhart, & Winston, Toronto,
- Grofman Bernard, Guillermo Owen, and Scott L. Feld (1983), Thirteen theorems in search of the truth”, *Theory and Decision*, 15, 261–278.
- Hampton, J.M., P.G. Moore, and H. Thomas (1973), Subjective probability and its measurement, *Journal of the Royal Statistical Society*, 136, part 1, 21-42.
- Havrda, J. and F. Chevrat (1967), Quantification method of classification processes: The concept of structural α - entropy, *Kybernetika*, 3, 30-35.
- Johnstone, David, J. (2007), The Value of a Probability Forecast from Portfolio Theory, *Journal Theory and Decision*, 63(2: Sept) , 153-203.
- Jøsang, Audun (2001), A Logic for Uncertain Probabilities, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3:June), 279–311
- Jøsang, Audun and Rosland Ismail (2002), The beta reputation system, *15th Bled Electronic Commerce Conference e-Reality: Constructing the e-Economy*, Bled, Slovenia, June 17-19.

- Jose, Victor Richmond R. , Robert F. Nau, and Robert L. Winkler (2007), Scoring Rules, Entropy, and Imprecise Probabilities, *5th Symposium on Imprecise Probability: Theories and Applications*, Prague, Czech Republic.
- Jose, Victor Richmond R., Robert F. Nau, and Robert L. Winkler (2008), Scoring Rules, Generalized Entropy, and Utility Maximization, *Operations Research*, forthcoming.
- Keynes, John M. (1921), *A Treatise on Probability*, London, Macmillan.
- Kleiter, Gernot D. (1996), Propagating imprecise probabilities in Bayesian networks, *Artificial Intelligence*, 88, (1-2: December), 143-161.
- Kleiter, Gernot D. (1994) Expressing imprecision in probabilistic knowledge, *Journal of the Italian Statistical Society*, 2, 213-223.
- Kullback, Solomon (1997), *Information Theory and Statistics*, Mineola, New York, Courier Dover.
- Kullback, Solomon (1959), *Information Theory and Statistics*, New York, Wiley.
- Ladha, Krishna, K. (1992), The Condorcet Jury Theorem, Free Speech, Correlated Votes, *American Journal of Political Science*, 3(36), 617-634.
- List, Christian and Robert E. Goodin (2001), Epistemic democracy: Generalizing the Condorcet jury Theorem, *Journal of Political Philosophy*, 9(3), 277–306.
- Lorge, I. and D. Fox, J. Davitz, and M. Brenner (1958), A survey of studies contrasting the quality of group performance and individual performance, 1920-1957, *Psychological Bulletin*, 5(6:November),337-372.
- Morris, Peter A. (1986), Observations on Expert Aggregation, *Management Science*, 32(3), 321-328.
- Morris, Peter A. (1983), An axiomatic approach to expert resolution, *Management Science*, 29, 24-32
- Press, S. James (1989), *Bayesian statistics: principles, models, and applications*, New York, Wiley.
- Savage, Leonard J. (1972), *The Foundations of Statistics*. Dover Publications, New York, second edition
- Savage, Leonard J. (1971) Elicitation of Personal Probabilities and Expectations, *Journal of the American Statistical Association*, 66(336), 783-801.
- Savage, Leonard J. (1954), *The Foundations of Statistics*, New York, Wiley.

Selten, Reinhard (1998), Axiomatic Characterization of the Quadratic Scoring Rule, *Experimental Economics*, 1,43-62.

Sunstein, Cass R. (2006), *Infotopia: How Many Minds Produce Knowledge*, Oxford, Oxford University Press.

Sunstein, Cass R. (2005), Deliberation and Information Markets, *Information Markets: A New of Making Decisions*, Robert W. Hahn and Paul C. Tetlock, editors, Washington D.C., AEI Brookings, 67-100.

Surowiecki, James (2004), *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, New York, Doubleday.

Toubia, Oliver (2006), Idea Generation, Creativity, and Incentives, *Marketing Science*, 25(5: Sept-Oct), 411-425.

Toubia, Oliver and Laurent Florès (2007), Adaptive Screening of Ideas, *Marketing Science*, 26(3: May-June), 342-363.

Winkler, Robert L. (1986), Expert resolution, *Management Science*, 32, 298-303.

Winkler, R. L. (1977), Rewarding expertise in probability assessment, 127-140. H. Jungermann and G. de Zeeuw, eds., *Decision-Making and Change in Human Affairs*, Dordrecht, Holland: D. Reidel, 45-56.

Winkler, Robert L. (1968), The consensus of subjective probability distributions, *Management Science*, 15 (2), B61-B75.

Winkler, Robert L. (1967), The quantification of judgment: Some methodological suggestions, *Journal of the American Statistical Association*, 62,1105-1120.