Scoring Rules for Competing Predictions

Willem K.B. Hofstee & Klaas Nevels

University of Groningen, The Netherlands

S A M E N V A T T I N G:

Scoringsregels om vast te stellen welke van twee partijen meer gelijk krijgt bij verschil in predictie in een binair experiment worden gepresenteerd. Eigenschappen van deze scoringsregels worden gecontroleerd. Vervolgens wordt uiteengezet hoe een methodologie waarin uitspraken worden opgevat als weddenschapsaanbod zich tot andere statistische modellen verhoudt.

Trefwoorden: scoringsregels, statistiek, weddenschap, methodologie, predictie, subjectieve kans.

Paper presented at the European Meeting of the Psychometric Society, Groningen, 1980

Imagine that you would want to make a bet on the outcome of the next presidential election. Suppose your prediction is that Mr. C. will win after all. Much to your dissappointment you are unable to trace any creditable persons who disagree with you to the point of venturing even a bottle of California wine. Should you give up and look for other experiments, natural or contrived, to satisfy your need for being put in the right?

Not necessarily. Under the circumstances, the personal predictive probability that Mr. C. will win is apparently $>.5$ for each potential participant, assuming a binary outcome of the experiment. But that does not mean that there is no difference in expectation. You should be able to find others who are more, or who are less confident that the outcome will be C. In fact, your chances of finding someone whose predictive probability is identical to yours are zero.

So all you need is a satisfactory rule that tells you how much a participant receives from another if, for example, his or her probability of C is .7 and the other person's p(C) = .9, and if indeed C occurs.

Once such rule is the quadratic rule:

$$v_{AB} = q_B^2 - q_A^2 ,$$

in which:

v_{AB} = the payment, in arbitrary units, that participant A receives from B;

q_A = A's predictive probability that the actual outcome of the binary experiment would not occur.

The payoff-matrix for the above example is given in Table 1. The quadratic rule may be found satisfactory since it meets all

Table 1.

Payoff-matrix for a binary bet under the quadratic scoring rule: $p_A(C) = .7$ and $p_B(C) = .9$.

	Outcome	
participant	C	Not C
A	$.1^2 - .3^2 = -.08$	$.9^2 - .7^2 = .32$
B	$.3^2 - .1^2 = .08$	$.7^2 - .9^2 = -.32$

of the following requirements:

R1. Reproducingness. Each participant maximizes its expected payoff by submitting a predictive probability which is a veridical representation of its subjective beliefs. So the rule forces a rational participant to say what he or she thinks. In the following, it will be assumed that a participant behaves accordingly.

R2. Positive expected payoff. Although the bet is a zerosum game under the quadratic rule, the subjectively expected payoff is positive for both participants.

R3. Symmetry. The subjectively expected payoff is equal for the two participants.

R4. Boundedness. The maximum amount that can be lost or won is unity.

Appendix 1 proves that the quadratic rule is the only rule that satisfies all the requirements R1 through R4. Another interesting rule is the logarithmic rule:

$$V_{AB} = {}^2\log(p_A/p_B)$$

The payoff-matrix for the election example is given in Table 2. The log rule does not satisfy R3 en R4, but it may be found

Table 2.

Payoff-matrix for a binary bet under the logarithmic scoring rule: $p_A(C) = .7$ and $p_B(C) = .9$

	Outcome	
participant	C	Not C
A	${}^2\log {}^7/_9 = -.36$	${}^2\log {}^3/_1 = 1.58$
B	${}^2\log {}^9/_7 = .36$	${}^2\log {}^1/_3 = -1.58$

useful precisely because the maximum amount that can be lost is unbounded. If a participant is subjectively certain (p=1) about an event that does not in fact occur, it may be judged that no punishment is heavy enough.

Generalizations of the quadratic and logarithmic rules for more than two possible outcomes will be presented elsewhere (Hofstee, 1980). Since a dependent variable can always be dichotomized, the present rules will probably meet the most urgent demands on the part of betting addicts.

In the remainder of this paper, this kind of betting will be taken as a methodological paradigm (Hofstee, 1977), to be compared with other methodological models. The central characteristics of the betting model are the following:

(1) It reflects a 'predictivist' conception of empirical science, as opposed to an inductivist conception. The criteria for an empirical statement are forever located in a future state of affairs, and no evidence from the past is able to serve as a justification of a prediction. Thus the betting model is in line with classical statistics as opposed to bayesian statistics. As in classical statistics , the appropriate form of an empirical statement under the betting model is a conditional or unconditional predictive distribution $p(x|\theta)$ or $p(x)$, not a posterior distribution $p(\theta|x)$. Other methodological models to be discussed here all share this predictivist character; although liberal use is made of prior distributions in some of these models, they are not bayesian in the inductivist sense.

(2) Clearly, 'betting model' implies that at least two competing hypotheses are at stake rather than just one.

(3) Vague hypotheses are acceptable under the betting model. Thus the uncertainty of prediction may arise not only from finite sample sizes, but also from personal uncertainty about the location of the parameter.

(4) A final characteristic of the betting model is a reproducing scoring or desicion rule, as illustrated above.

Taking the characteristics (2), (3) and (4) as variables, a taxonomy of predictivist methodologies results which is depicted in Table 3. The cells of that table will be commented

Table 3

Taxonomy of 'predictivistic' methodologies

	single hypothesis		competing hypotheses	
	nonreproducing scoring rule	reproducing scoring rule	nonreproducing scoring rule	reproducing scoring rule
point prior	1 null hypothesis testing		4 e.g., Koele (1979)	
prior distribution	2 generalized testing Model	3 Van Naerssen 1961	5 Pitz (1978)	6 betting model: Hofstee e.a., 1980

upon in backward order.

Model 5: Pitz (1978) presented a methodological paradigm in which two competing predictive distributions of a future observation are derived from two prior distributions (which Pitz called 'prediction distributions', may be for fear of being decried as a bayesian), and in which the support of one hypothesis over another is measured by the ratio

$$\frac{p_A(x_j)}{p_B(x_j)}$$

of the two likelihoods of the observed outcome under the two predictive distributions. Since Pitz explicitly mentions the possibility of betting, the probabilities may be conceived as odds, and therefore his payoff-rule may be constructed to be the following:

$$V_{AB} = \frac{p_a(x_j)}{p_A(x_j)+p_B(x_j)}$$

This model is essentially identical to our 1977 version (Hofstee, 1977) of the betting model. Appendix 2, however, proves that the ratio rule is not reproducing. This lack of reproducingness is an undesirable feature since the ratio-rule

would force a rational participant to submit a prior distribution that deviates in a systematic manner from his or her own beliefs. It is hard to see how a rule which forces that kind of strategic behavior can be justified methodologically.

Model 4: Comparative testing of two exact or one-sided hypotheses is the topic of classical statistical decision theory. In this tradition, Koele (1979) has recently proposed a procedure in which a minimal sample size is calculated such that the critical regions of the two hypotheses do not overlap; consequently, one hypothesis can be rejected in favor of the other hypothesis. Since the procedure is two steps away from the betting model, its drawbacks are twofold from that point of view. In the first place, the decision rule is not reproducing : it may be intuitively clear that a participant can increase his or her chances of being put in the right by moving closer to the opponents' hypothesis (for a formal proof, see Appendix 3). Since this holds for both participants, they would move closer and closer and in the end only an infinite sample could solve their difference of opinion. In the second place, it is difficult to see the use of exact or one-sided hypotheses. Participant who would express themselves in terms of such hypotheses can only be people who are infinitely prejudiced in the sense that they award prior probabilities of zero to certain intervals of the parameter scale.

Model 3: Van Naerssens (1961) classical article on the scaling of subjective probability contains reproducing scoring rules for single predictive distributions in a binary experiment. Apart from certain refinements, this methodological model is only one step away from the betting model. Its only drawback is the single hypothesis character. The model cannot distinguish between events that are easy versus difficult to predict. Therefore under the model, the rational investigator would be forced to look for areas in which success in virtually assured. Under the betting model, the other person's predictive ability forms a natural baseline. Therefore the methodological moral is quite different.

Model 2 may be constructed as the case in which a vague rather than an exact hypothesis is tested according to the classical statistical procedure. The objections to both this model and Model 1 have

been put forward by implication.

A final comment should be made on the methodological philosophy that underlies the betting model. Contrary to the conception of empirical science as a serene and disinterested search for truth, the present approach views science as a very human affair in which individual investigators work on their reputation. A methodology should then be a set of rules and sanctions that forces people to do the right thing if they wish to maximize their expected reputation. Doing the right thing might mean, for example, not to engage in selective publication and not to produce truisms through rejecting exact hypotheses that no sensible person would ever believe in anyway. There is a fair amount of consensus that the null hypothesis testing methodology forces the rational investigator to display just such undesirable behavior. The betting model may be viewed as an attempt to correct for the corrupting influence of the null hypothesis testing methodology, while retaining its most fundamental characteristic, i.e.,the predictivist command that hypotheses should be stated in advance.

R E F E R E N C E S:

Hofstee, W.K.B. De weddenschap als methodologisch model. Nederlands Tijdschrift van de Psychologie, 1977, 32, 203 - 217.

Hofstee, W.K.B. De empirische discussie: Theorie van het sociaal-wetenschappelijk onderzoek. Meppel: Boom, 1980.

Koele, P. Afscheid van de nulhypothese. Kennis en Methode, 1979, 3, 446-458.

Naerssen, R.F. van, A scale for the measurement of subjective probability. Acta Psychologica, 1961, 17, 159-166.

Pitz, G.F. Hypothesis testing and the comparison of imprecise hypotheses. Psychological Bulletin, 1978, 85, 794-809.

APPENDIX 1.

We first show that the quadratic rule satisfies the requirements R1 through R4.

R1. Let p_A^* be the probability submitted by A. It is sufficient to show that

(1) $(q_B^2 - q_A^{*2})p_A + (p_B^2 - p_A^{*2})q_A \leq (q_B^2 - q_A^2)p_A + (p_B^2 - p_A^2)q_A$.

For the difference between the left and right side of the inequality in (1) we find

(2) $(q_A^2 - q_A^{*2})p_A + (p_A^2 - p_A^{*2})q_A =$

$$= \left\{(1 - p_A)^2 - (1 - p_A^*)^2\right\}p_A + (p_A^2 - p_A^{*2})(1 - p_A)$$

$$= (p_A^* - p_A)\left\{(2 - p_A - p_A^*)p_A - (p_A + p_A^*)(1 - p_A)\right\}$$

$$= -(p_A^* - p_A)^2 \leq 0.$$

Remark that equality in (1) is attained if and only if $p_A^* = p_A$.

R2. The subjectively expected payoff for e.g. A equals

(3) $(q_B^2 - q_A^2)p_A + (p_B^2 - p_A^2)q_A = \left\{(1 - p_B)^2 - (1 - p_A)^2\right\}p_A +$

$$(p_B^2 - p_A^2)(1 - p_A)$$

$$= (p_A - p_B)\left\{(2 - p_A - p_B)p_A - (p_A + p_B)(1 - p_A)\right\}$$

$$= (p_A - p_B)^2 > 0.$$

R3. See the result derived in (3).

R4. Trivial.

In what follows we show that for dichotomous situations the quadratic rule is the only rule that satisfies all the requirements R1 through R4. The two possible outcomes of the binary experiment are called success and failure. Let A and B be the players, with p_A (resp. p_B) the predictive probability of A (resp. B) that success occurs and r_A (resp. r_B) the probability which A (resp. B) assigns to the success.

Let $v(r_A, r_B)$ denote the payment that A receives from B if the success occurs. With relation to the function v we assume that v is continuous at each point (x,y) of the rectangle $R = \{(x,y),\ 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$ and differentiable in the interior of R. We remark that the boundedness of v follows from the required continuity of v on R (requirements R_4). Moreover we assume that the function v satisfies the following obvious properties:

$$
\begin{aligned}
&\text{(a)} \quad v(x,y) < 0 \ (>0) \quad \text{if } x < y \ (x > y) \\
&\text{(b)} \quad v(x,x) = 0 \\
&\text{(c)} \quad v(x,y) = -v(y,x).
\end{aligned}
\tag{4}
$$

If r = p for both players, then the subjectively expected value of A (resp. B), notation SEV_A (resp. SEV_B) equals

$$
\begin{aligned}
&\text{(a)} \quad SEV_A = p_A\, v(p_A, p_B) + (1 - p_A)\, v(1 - p_A, 1 - p_B), \\
&\text{(b)} \quad SEV_B = p_B\, v(p_B, p_A) + (1 - p_B)\, v(1 - p_B, 1 - p_A).
\end{aligned}
\tag{5}
$$

In order that R_3 holds it is necessary and sufficient that

$$
v(1 - p_A, 1 - p_B) = \frac{p_A + p_B}{p_A + p_B - 2}\, v(p_A, p_B), \quad p_A + p_B < 2. \tag{6}
$$

Relation (6) follows from (5) and property 14.c.
Combining (5) and (6) gives

$$
SEV_A = SEV_B = \frac{p_A - p_B}{2 - p_A - p_B}\, v(p_A, p_B), \quad p_A + p_B < 2. \tag{7}
$$

Remark that if $p_A + p_B = 2$, then $SEV_A = SEV_B = 0$.
From (7) it follows with (4) that for both players the subjectively expected payoff is non-negative (requirement R2). The function w, defined by

$$
v(x, y) = (x - y)\,(2 - x - y)\, w(x, y), \quad x \neq y \tag{8}
$$

is for $x \neq y$ uniquely determined by v.
From (7) now it follows with (8)

$$
SEV_A = SEV_B = (p_A - p_B)^2\, w(p_A, p_B). \tag{9}
$$

From the definition of w and the properties of v it follows that $w(x,y) > 0$ for $x \neq y$.

We now prove that if R3 holds, then a necessary condition that R1 holds is that the function w is positive and constant on the rectangle R. In general

$$(10) \qquad SEV_A = (r_A - r_B)(2p_A - r_A - r_B)\, w(r_A, r_B)$$

In order that the right side of (10) achieves its maximum for $r_A = p_A$ with $0 < p_A < 1$, it is necessary that

$$(11) \qquad \frac{\delta}{\delta r_A}\Big[(r_A - r_B)(2p_A - r_A - r_B)\, w(r_A, r_B)\Big]_{r_A = p_A} = 0.$$

Carrying out the differentiation in (11), we find that

$$(12) \qquad \frac{\delta w(p_A, r_B)}{\delta p_A} = 0 \quad \text{for all } 0 < p_A < 1.$$

In the same way we obtain

$$(13) \qquad \frac{\delta w(r_A, p_B)}{\delta p_B} = 0 \quad , \quad 0 < p_B < 1.$$

From (12) and (13) and the continuity of v on the rectangle R now it follows that

$$(14) \qquad w(x,y) = k > 0 \qquad \text{on R.}$$

Combining (8) and (14) gives for $k = 1$

$$(15) \qquad \begin{aligned} v(p_A, p_B) &= (p_A - p_B)(2 - p_A - p_B) \\ &= (p_A - 1 + 1 - p_B)(1 - p_A + 1 - p_B) \\ &= (q_B - q_A)(q_A + q_B) \\ &= (q_B^2 - q_A^2). \end{aligned}$$

APPENDIX 2

We shall prove here that the scoring rule fitting the ratio rule is not reproducing, by showing that this scoring rule satisfies the requirements R2 through R4.

From this and the result derived in appendix 1 it then follows that this rule does not satisfy R1. For simplicity we only give the proof for dichotomous situations. The same notation as in appendix 1 is used. So if the success occurs, then A receives $r_A/(r_A + r_B)$ of the total stake S. It is no restriction to take S = 1.

If r = p for both players, then

$$(16)\quad \text{(a)}\quad SEV_A = \frac{p_A^2}{p_A+p_B} + \frac{q_A^2}{q_A+q_B} > 0,$$

$$\text{(b)}\quad SEV_B = \frac{p_B^2}{p_A+p_B} + \frac{q_B^2}{q_A+q_B} > 0.$$

From (16) it follows that $SEV_A = SEV_B$, since

$$(17)\quad SEV_A - SEV_B = \frac{p_A^2 - p_B^2}{p_A + p_B} + \frac{q_A^2 - q_B^2}{q_A + q_B} = p_A - p_B + q_A - q_B = 0.$$

Now we know that this scoring rule does not satisfy R1.

Let us assume that player A wishes to submit a probability r_A so as to maximize SEV_A. Therefore we consider the function

$$(18)\quad f(r) = \frac{r}{r + p_B} p_A + \frac{1 - r}{1 - r - q_B} q_A$$

and investigate for which value of r the function f(r) achieves its maximum on the closed interval $[0,1]$.

Since

$$(19)\quad f'(r) = \frac{p_a p_B}{(r+p_B)^2} - \frac{q_A q_B}{(1-r+q_B)^2},$$

we find that if $-p_B < r < 1 + q_B$, then $f'(r) = 0$ for $r = c_o$, where

$$(20)\quad c_o = 2\,\frac{\sqrt{p_A p_B}}{\sqrt{p_A p_B}+\sqrt{q_A q_B}} - p_B .$$

Let r^* with $0 \leq r^* \leq 1$ be the value of r for which the function f(r) achieves its maximum.

Then from the above it follows that

if $c_o < 0$, then $r^* = 0$,

if $0 \leq c_o < 1$, then $r^* = c_o$,

if $c_o > 1$, then $r^* = 1$.

(for an illustration, see Figure 1)

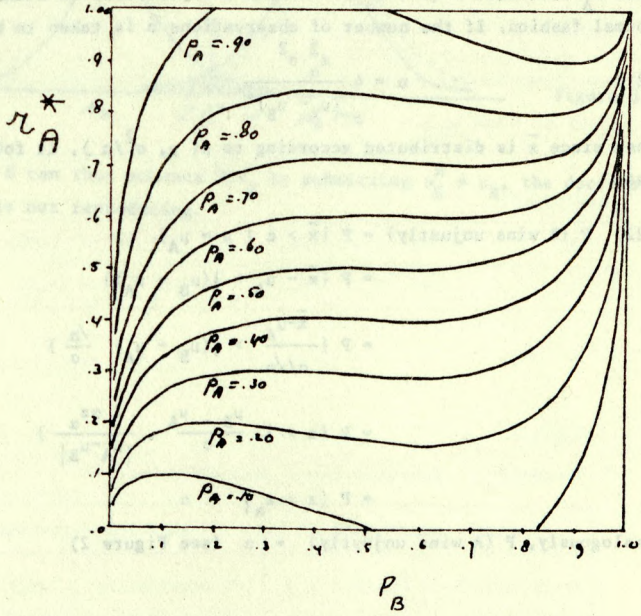


Figure 1.

APPENDIX 3

Players A and B have a difference of opinion regarding a certain parameter μ . According to A, $\mu = \mu_A$ and according to B, $\mu = \mu_B$. Without loss of generality we take $\mu_A < \mu_B$. Both players stake an equal amount, the sum of which is taken to be 1. A number of observations are made. The players agree that the observations are sampled from a normal distribution with mean μ and known variance σ^2. If the mean $\bar{x}$ of the observed values is less than $c = \frac{1}{2}(\mu_A + \mu_B)$, A receives the total stakes. If $\bar{x} > c$, B receives the total stakes. If $\bar{x} = c$, the bet is undecided. A further requirement is that the probability α that A wins unjustly is equal to the probability that B wins unjustly.

Take z_A such that $P(z > z_\alpha) = \alpha$ with z distributed in standard normal fashion. If the number of observations n is taken to be

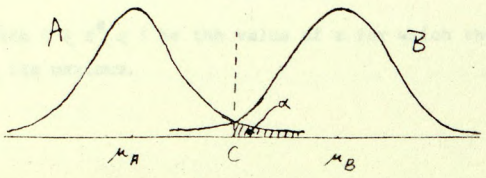
$$n = 4 \frac{z_\alpha^2 \sigma^2}{(\mu_A - \mu_B)^2} \quad , \tag{21}$$

then since $\bar{x}$ is distributed according to $\mu(\mu, \sigma^2/n)$, it follows that

$$\begin{aligned} (22)\quad P\{\text{B wins unjustly}\} &= P\{\bar{x} > c \mid \mu = \mu_A\} \\ &= P\{\bar{x} - \mu_A > \tfrac{1}{2}(\mu_B - \mu_A)\} \\ &= P\left\{\frac{\bar{x}-\mu_A}{\sigma/\sqrt{n}} > \tfrac{1}{2}(\mu_B - \mu_A)\ \frac{\sqrt{n}}{\sigma}\right\} \\ &= P\left\{z > \tfrac{1}{2}\ \frac{\mu_B - \mu_A}{\sigma} \cdot \frac{2\sigma z_\alpha}{|\mu_A - \mu_B|}\right\} \\ &= P\{z > z_\alpha\} = \alpha \end{aligned}$$

Analogously, P {A wins unjustly} = α (see Figure 2)

Figure 2.



It is further seen that

(23) $\quad SEV_A = SEV_B = 1 - 2\alpha > 0$

Now consider the case in which B does not submit his true μ_B, but submits $\mu = \mu_B^*$ such that $\mu_A < \mu_B^* < \mu_B$. Consequently, $c^* = \frac{1}{2}(\mu_A + \mu_B^*)$ and $n^* = 4 z_\alpha^2 \sigma^2 / (\mu_A - \mu_B^*)^2 > n$.
As a result, $SEV_A = 1 - 2\alpha$, while $SEV_B = 1 - 2\alpha^* > SEV_A$, since $\alpha^* < \alpha$ (see Figure 3).

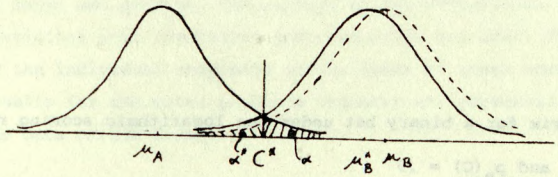


Figure 3.

Since B can thus enhance SEV_B by submitting $\mu_B^* \neq \mu_B$, the decision rule is not reproducing.

Table 1.

Payoff-matrix for a binary bet under the quadratic scoring rule:

$p_A(C) = .7$ and $p_B(C) = .9$.

	Outcome	
participant	C	Not C
A	$.1^2 - .3^2 = -.08$	$.9^2 - .7^2 = .32$
B	$.3^2 - .1^2 = .08$	$.7^2 - .9^2 = -.32$

Table 2.

Payoff-matrix for a binary bet under the logarithmic scoring rule:

$p_A(C) = .7$ and $p_B(C) = .9$

	Outcome	
participant	C	Not C
A	$^2\log {}^7/_9 = -.36$	$^2\log {}^3/_1 = 1.58$
B	$^2\log {}^9/_7 = .36$	$^2\log {}^1/_3 = -1.58$

Table 3

Taxonomy of 'predictivistic' methodologies

	single hypothesis		competing hypotheses	
	nonreproducing scoring rule	reproducing scoring rule	nonreproducing scoring rule	reproducing scoring rule
point prior	1 null hypothesis testing		4 e.g., Kaele (1979)	
prior distribution	2 generalized testing Model	3 Van Naerssen 1961	5 Pitz (1978)	6 betting model: Hofstee e.a., 1980