The logarithmic scoring rule of decision theory

1. How do you grade a weatherman?

Weatherman A says there's a 30% chance of rain Monday. Weatherman B disagrees; he says there is a 45% chance. Say it rains Monday. Based on this much evidence, we should favor the predictions of Weatherman B. Suppose next that Weatherman A says there is a 65% chance of rain Tuesday; Weatherman B says 80%. It doesn't rain on Tuesday. Now who looks to be the better weatherman? To answer this, we need a way to keep score.

A naive strategy would be to give the weatherman a score equal to the probability they ascribe to the actual outcome. So Weatherman A would get .3 points for his Monday prediction and .35 points for his Tuesday prediction, while Weatherman B would get .45 points for his Monday prediction and .2 points for his Tuesday prediction. Both would be at .65 points–a tie.

This turns out not to be a very good way to score weathermen. For, suppose we think the probability of rain is 70%. If we say 70%, we score .7 if it rains and .3 if it doesn't. So our *expected score* (according to our own estimate) is (.7)(.7)+(.3)(.3)= .58. We'd do better to just say 100%, when our expected score would be (.7)(1.0)+(.3)(0)=.7. The upshot is that a weatherman who knows the rule will only give predictions of 0% and 100%–not a desired feature.

Another idea would be to prefer the weatherman who assigns the highest probability to the complete set of observations {rain Monday, no rain Tuesday}. Since (presumably) the Tuesday prediction was conditioned on the observation of rain on Monday, the conditional probability law $P(E \cap F) = P(E)P(F|E)$ tells us that Weatherman A assigned probability (.3)(.35)=1.05 and Weatherman B assigned probability (.45)(.2)=.9 to that set of observations. According to this analysis (which will turn out to be essentially correct), Weatherman A's performance is best.

2. Logarithmic scoring.

A "scoring rule" for weathermen (and prognosticators generally) is implemented as follows: choose some non-decreasing function f on $(0, 1)$. Suppose a weatherman assigns a probability q to some event (rain Monday, for example). Assign a score $f(q)$ if the event occurs and $f(1-q)$ if it doesn't. But what function $f(x)$ should be chosen? Various scoring rules exist, but $f(x) = \log x$ is of special interest. (A *logarithmic* scoring rule.)

In favor of the logarithmic scoring rule is the fact that the expected score it assigns is $S(q) = p \log q + (1-p)\log(1-q)$, where p is the *actual* probability of rain, which is maximized at $q = p$. To see this, we calculate $\frac{d}{dq}S(q) = \frac{p}{q} + \frac{p-1}{1-q}$. (We assume here that these are natural logarithms, though any base will do.) This is zero at $q = p$, and one may check that this critical point corresponds to the global maximum of S.

On the other hand, there are many functions $f(x)$ having the property that the global maximum of $pf(q) + (1-p)f(1-q)$ occurs at $q = p$. A more convincing argument for $f(x) = \log x$ comes from considering a situation in which there are more than two possible outcomes. (This was considered in [1].) Consider for example the case of an unfair die with n faces. A predictor picks positive numbers $q_1, q_2, \ldots, q_n$, with $q_1 + q_2 + \cdots + q_n = 1$

to represent his estimation of the respective probabilities of these faces landing down upon a toss of the die. The player scores $f(q_i)$ in the event of the ith outcome.

Taking $n = 3$, the expected score is $S(q_1, q_2) = p_1 f(q_1) + p_2 f(q_2) + (1 - p_1 - p_2) f(1 - q_1 - q_2)$, where p_1 and p_2 are the actual probabilities of outcomes 1 and 2. The function S has a critical point where its partial derivatives $S_{q_1} = p_1 f'(q_1) - (1 - p_1 - p_2) f'(1 - q_1 - q_2)$ and $S_{q_2} = p_2 f'(q_2) - (1 - p_1 - p_2) f'(1 - q_1 - q_2)$ are both equal to zero. If this critical point corresponds to the global maximum of S, and we want this to occur at $q_1 = p_1$, $q_2 = p_2$, this yields $p_1 f'(p_1) = p_2 f'(p_2)$. But this should hold for all p_1, p_2 with $0 < p_1 + p_2 < 1$. In otherwords, $x f'(x)$ is constant on $(0, 1)$ and one quickly determines $f(x)$ to be some constant multiple of $\log x$.

Yet another argument for the logarithmic scoring rule comes from consideration of the *surprisal* of an observation, which is basically a numerical measure of how "surprised" a prognosticator is by the observation we are considering...or to put it somewhat differently, a measure of how much information is gained by way of the observation.

For example, upon learning the result of a fair coin toss, one gains 1 bit of information, which is $-\log_2 \frac{1}{2}$; note that $\frac{1}{2}$ is the probability previously assigned to the observed result (be it heads or tails). More generally, upon making any observation, the result of which was previously assigned probability q, the surprisal consists in $-\log_2 q$ bits of information. A progonisticator is doing a good job prognosticating precisely to the degree that he is able to minimize surprisal (the very best prognosticator would not be at all surprised by anything), and of course to minimize the expectation of the surprisal $-\log_2 q$ is precisely to maximize expected performance under the logarithmic scoring rule.

3. Your office football pool.

So how exactly should you score your office football pool? When I was in graduate school, my fellow math TAs were representatively unimaginative: everyone paid a dollar and picked a winner for each NFL game–who picked the most correctly got the money. One of us (John) made his picks by tossing a coin, and did surprisingly well over time. Surely, I figured, there must be some system that would favor excellence in prognostication more consistently. I wasn't a big fan of betting against a point spread, though, because once the outcome of a game is sealed, teams aren't generally competing at peak effort.

Could the logarithmic scoring rule provide what I was looking for? Here's the general idea: the players submit probabilities for each game, are scored according to the rule, and it's winner take all as before. Making the predictions becomes more interesting, scoring a bit more tedious. Note, however, you don't actually have to take logarithms; $\log p_1 + \cdots + \log p_n > \log q_1 + \cdots + \log q_n$ if and only if $p_1 \cdots p_n > q_1 \cdots q_n$, so one can choose to multiply the appropriate probabilities rather than add their logarithms.

Does the system produce the desired result of rewarding excellence more consistently? This isn't an easy question to answer–certainly not always. Consider for example the case of John the coin tosser playing against his sister Janet. Janet doesn't watch football either, but has read that home teams win 57% of the time in the NFL. John, presumably, calls each game at probability .5 for the home team; Janet says .57. Taking logs base 2 and adding 1 to each player's score for normalization, John scores $\log_2(.5) + 1 = 0$ regardless of who wins, while Janet scores $\log_2(.57) + 1 \approx .189034$ for a home win and

$\log_2(.43) + 1 \approx -.21759$ for a win by the visitors. This is not good for Janet; if 8 home teams win and 8 lose, John takes her money. In fact it could be much worse. If only the 15 Sunday games count and 8 home teams win, John *still* takes her money! Janet retains a slight edge (she wins with probability $\sum_{i=9}^{15}(.57)^i(.43)^{15-i}\binom{15}{i} \approx .515$), but she'd do far better if they were just picking winners.

The philosophy supporting the logarithmic scoring rule is so solid, however, that it's worth looking at a more realistic (if still artificial) example involving several (10) players having various states of knowledge. Here is such a model: "games" are contests between two teams of nine fair dice each having various numbers of faces. High pip count from a single round of throws wins (ties broken randomly). The strength of a team of dice is determined by the number of faces on its constituents. For each die, this number is itself determined by a prior roll of a fair six-sided die. The ten players competing in the pool have graded amounts of knowledge; Player N knows how many faces are on each team's first $N-1$ dice. Teams play just once, then retire.

For example, say Player 3 knows that Team A's first two dice have 5 and 4 sides respectively, while Team B's first two dice have 2 and 3 sides. He estimates the probability that A wins the contest using the normal approximation, as follows. Let $S = U_5 + U_4 - U_3 - U_2 + \sum_{i=1}^{7}(D_i - C_i)$, where U_n are independent uniform variables on $\{1, \ldots, n\}$ and each of C_i, D_i are independent variables distributed as U_{U_6} on $\{1, 2, \ldots, 6\}$ (where an initial roll of a 6 sided die determines the n of the subsequently evaluated variable U_n). Team A wins if $S > 0$; if $S = 0$ a coin is tossed to determine the winner. The expectation of S is $E(U_5)+E(U_4)-E(U_3)-E(U_2) = 3+2.5-2-1.5 = 2$, as $E(D_i - C_i) = 0$ for all i, while the variance is $\frac{1}{4}+\frac{2}{3}+\frac{5}{4}+2+14\cdot\frac{275}{144} \approx 30.903$ for a standard deviation of around $\sqrt{30.903} \approx 5.56$. The normal approximation thus yields $P(\text{A wins}) \approx P(Z < \frac{2}{5.56}) \approx N(.36) \approx .64$. (In order for B to win, S must fall at least .36 standard deviations left of its mean.)

A computer simulation of 14,000 16-game pools among these 10 contestants gave the following victory rates:

	Picking winners	Logarithmic scoring
Player 1	1.83%	2.59%
Player 2	4.56%	2.23%
Player 3	5.37%	2.37%
Player 4	6.45%	2.96%
Player 5	7.20%	3.90%
Player 6	8.31%	4.87%
Player 7	9.89%	6.65%
Player 8	12.68%	9.42%
Player 9	17.18%	17.30%
Player 10	26.53%	47.71%

Here, logarithmic scoring seemed to serve the most knowledgable player quite well.

4. Guessing games.

Here's one to try. For your next game of twenty questions, have the guesser include his credence q in a "yes" response with every question, and, instead of charging a single question each time, charge fractional (real valued) amounts based on the logarithmic scoring

rule, i.e. $-\log_2 q$ if "yes" and $-\log_2(1-q)$ if "no". The guesser can always go with $q = .5$ and be charged 1 question. Or, they can indulge their desire to guess "Wilt Chamberlain" in the very first round (with $q = 2^{-20}$) without feeling like they are wasting their guess.

In 1972 at the University of Florida, students were given examinations with true-false and multiple choice questions. They were told to answer them probabilistically and were scored logarithmically (among other ways...see [2]). Be careful...if you assign zero probability to a correct answer, you score $-\infty$! No quantity of all-nighters will dig you out of that hole.

Anyone for *Clue*?

5. Sleeping Beauty.

Consider Beauty, whose case is a hot topic in contemporary analytic philosophy. She is involved in an experiment in which a fair coin is flipped on Sunday night. Beauty doesn't see the result of the toss, but surely her credence in *heads* is $\frac{1}{2}$ on Sunday night. She goes to sleep and is awakened Monday morning and asked for her credence again, after which she learns the true outcome. Now if (and only if) the outcome was tails, she receives a drug that puts her to sleep for another 24 hours and erases all memory of her Monday awakening. On Tuesday she is again awakened and asked for her credence in heads.

Beauty knows how the experiment works. If heads, she's awakened once; if tails, twice. She won't know for sure what day it is when she wakes up. The question is this: *when she wakes up Monday morning, what should her credence in heads be*? Some philosophers say $\frac{1}{2}$. They're called *halfers*. Others say $\frac{1}{3}$. They're called *thirders*.

What does the logarithmic scoring rule say? If Beauty assigns credence of q to *heads*, she'll score $\log q$ if in fact the coin landed heads. This happens with objective probability $\frac{1}{2}$. But if the coin landed tails, she will be asked for her credence twice (once on Monday and again on Tuesday). So maybe she should be scored twice. That makes her score $2\log(1-q)$ if the coin landed tails, which also happens with objective probability $\frac{1}{2}$. Her expected score is thus $S(q) = \frac{1}{2}\big(\log q + 2\log(1-q)\big)$. One has $\frac{d}{dq}S(q) = \frac{1}{2q} - \frac{1}{1-q}$. This is zero at $q = \frac{1}{3}$, and it is easily checked that S has a global maximum there. Apparently logarithmic scoring favors the one-third solution.

Or does it? Maybe Beauty has a "negative surprisal" when her memory is erased, and this should be scored as well. That would cancel one of the two "tails" surprisals, and we'd be back to the one-half solution. Which is correct? This is difficult to answer. The logarithmic scoring rule doesn't come with a user's manual. It's a powerful tool, but we must ultimately decide on its proper scope of application. There comes a point where mathematics ends, and philosophy begins....

References

[1] Shuford, Jr., Emir H., Arthur Albert and H. Edward Massengill. 1966. Admissible probability measurement procedures. *Psychometrika* **31(2)** 125-145.

[2] Irvin N. Glein and John B. Wallace, Jr., Probabilistically Answered Examinations: A Field Test, *The Accounting Review* **49** (1974), 363-366.