The RSS Crystal Ball 2010 scoring rule

Anthony O'Hagan

February 2, 2010

Abstract

This note explains the scoring rule that was chosen for the Royal Statistical Society's 'Crystal Ball' competition, 2010.

1 The scoring rule in brief

For each of the questions in the competition, contestants need to specify an estimate and a standard deviation to represent their judgements about an unknown quantity. After the deadline for competition entries, the true value of the quantity will become known for each question. The scoring rule then gives a score to each contestant for each question, and the contestant's total score is the sum of their scores on the various questions. The winner(s) will be decided on the basis of contestants' total scores, with low scores being better than high scores.

1.1 The rule ...

The rule for scoring a contestant for a given question depends on their estimate t, their standard deviation s, the true value x and a further value $\bar{s}$ (to be explained later). Specifically, the score has three components –

$$S = S_1 + S_2 + S_3 \ ,$$

where the components are

$$\begin{aligned} S_1 &= \left(\frac{x-t}{s}\right)^2 \ , \\ S_2 &= 2.5\sqrt{5}\exp\left\{-2\left(\frac{x-t}{s}\right)^2\right\} \ , \\ S_3 &= \left(\frac{x-t}{\bar{s}}\right)^2 \ . \end{aligned}$$

Although the first and third components are very similar, the difference between using s and $\bar{s}$ in the denominator is important.

The purpose of this note is to explain in detail how this scoring rule was chosen, why it was chosen in preference to other possible rules, and what the basic properties of this rule are. In principle, it is not really necessary for contestants to know the detail. Their task is simply to judge the uncertain quantity in each question as accurately as possible through their stated estimates, and to express the accuracy that they claim realistically through their stated standard deviations. However, this presupposes that the scoring rule is fair, and another purpose of this note is to reassure potential contestants on that matter.

1.2 ... may be fair ...

A fair scoring rule should reward (by giving low scores to) the contestants who perform the stated task well, i.e. to those who (a) are able to make good estimates, and (b) can assess realistically how good their estimates are. An unfair scoring rule would effectively reward people for doing something other than the stated task. If, for instance, better scores could be obtained by deliberately under-estimating the uncertain quantities, then this would not be a fair rule.

We should remember that luck inevitably plays some part in such a competition. Any one of the uncertain quantities could produce a really unexpected or odd value x. The best contestant might be beaten by someone luckier, who gave a poor estimate which just happened to coincide with a freak outcome. A fair rule will be one for which the contestant who best fulfils the stated task will get the best (lowest) scores on average, in the long run.

Our chosen rule is intended to be fair in this sense, and much of the detail in this note is about demonstrating this.

1.3 ... but ...

Having said that, the objective is not actually to get the best possible score on average, in the long run. The objective is to beat the other contestants. Suppose that you are a good judge of these quantities but you suspect you are not the best. You are going to need some luck to win, and to give rein to that luck you might choose to make riskier judgements. In what we have called a fair rule you will minimise your score in the long run by trying to carry out the task as specified, honestly and accurately specifying your estimates t and their standard deviations s, but it may be possible (through careful study of the rule) to increase your chances of actually winning the competition by doing something different (such as deliberately under-estimating the quantities).

Although it may be possible to gain an advantage in this way, remember that on average any such strategy will produce *worse* scores and we very much hope that contestants will enter into the spirit of the competition and enjoy tackling the task just as it is stated – give your best estimates and your best judgements (in the form of your standard deviations) of how good those estimates are.

Only readers actually interested in the technical detail need read further!

2 Another proper scoring rule

To introduce ideas for the scoring rule S, we first discuss the formal definition of what we have hitherto called a fair scoring rule, and then introduces a simpler rule.

2.1 Proper scoring rules

Suppose that you have a probability distribution for some unknown quantity X and are asked to give some summary p of that distribution. We distinguish between your actual value p and the value q that you state in response to being asked for p. After you have given the response q and the true value $X = x$ of the unknown quantity becomes available, you will receive a score $L(q, x)$. We will suppose that, as with our score S, lower scores are better. Before X becomes known, you do not know what score you will receive, but your expected score is $\bar{L}(q) = E\{L(q, X)\}$. The scoring rule is called proper if, no matter what your actual distribution may be for X, you will minimise your expected score by stating a value q that equals your actual value p, i.e.

$$\min_q \bar{L}(q) = \bar{L}(p) \ .$$

So a scoring rule is proper if you get the best score on average by honestly stating what is asked for. This is what we have called a fair rule in the previous section, but 'proper' is the usual technical term.

A simple example of a proper scoring rule is when p is the expected value of X, $p = E(X)$. Then consider the score $L(q, x) = (x - q)^2$. This is called the quadratic score, and it is easy to prove that it is proper. First note that

$$\begin{aligned} \bar{L}(q) &= E\{L(q, X)\} = E\left\{(X - q)^2\right\} = E(X^2) - 2qE(X) + q^2 \\ &= E(X^2) - 2qp + q^2 = (p - q)^2 + E(X^2) - p^2 \\ &= (p - q)^2 + Var(X) \ . \end{aligned}$$

This is obviously minimised by setting $p = q$.

More generally, if $p = E\{g(X)\}$ for some function $g(\cdot)$, then the scoring rule $L(q, x) = \{g(x) - q\}^2$ is proper.

2.2 Quadratic rule for the Crystal Ball competition

We can easily get a proper scoring rule now for the Crystal Ball competition by noting that we are asking for $t = E(X)$ and $s^2 = Var(X) = E(X^2) - t^2$. So we are equivalently asking for $E(X)$ and $E(X^2)$. So the following will be proper

$$L_q(t, s, x) = (x - t)^2 + (x^2 - s^2 - t^2)^2 \ .$$

The expected value of this score with respect to the contestant's distribution of X will be minimised by setting $t = E(X)$ and $s^2 + t^2 = E(X^2)$, which is the task we are setting them.

So if this is a proper scoring rule, why did we instead choose the more complicated score S?

The score has two components, one penalising poor assessment of t and the other penalising poor assessment of s. We have simply added the two components together above, but we could equally well have weighted them by a formula like

$$L_q^w(t, s, x) = w_t(x - t)^2 + w_s(x^2 - s^2 - t^2)^2 .$$

The score would clearly still be proper, but the weights allow us to give more or less emphasis to one component versus the other. This is important because the competition is primarily one of prediction and so t is somewhat more important than s. We would not want the winner to win because of accurately assessing uncertainty whilst being a poor estimator, so the second component should not dominate the score.

But it is hard not to have the second term dominating, because it effectively concerns the fourth power of x (or s or t), while the first term is only second power. It will be hard to balance the two for any given question. In general, a score that increases according to the fourth power as x becomes further from t is very stringent in the way it penalises poor assessment.

Furthermore, we want to balance the scores across the various questions, so that all have more or less equal importance in forming the contestant's total score. When the answers to the questions will be of quite different magnitudes, this will again be difficult. The problems of weighting the components of the score, both within and between questions, is the principal reason for looking for another scoring rule.

A lesser, but still real, concern is the asymmetry of this rule. Underestimating x by a given amount does not lead to the same penalty as overestimating it by the same amount.

3 Interpretation of S

We begin our study of the proposed scoring rule S by looking at its general behaviour in order to understand its three components, and to see why it is preferred to the quadratic score.

We can see that the third component is the quadratic score $(x-t)^2$ weighted by being divided by $\bar{s}^2$. The expectation of this component is minimised by setting $t = E(X)$, and one way to look at the role of $\bar{s}$ is simply as determining a weight for this third component.

The first component has the same formula except that the contestant's own standard deviation s replaces $\bar{s}$. It is also a quadratic score that is minimised on average by $t = E(X)$, but that is no longer its only purpose because it works in conjunction with the second component.

The second component, S_2, is more complex. Notice that it is just a function of $S_1 = (x - t)^2/s^2$, but it is actually a *decreasing* function of this quadratic function. So while S_1 increases with distance of x from t we find that S_2

decreases. The reason for this is to penalise poor specification of s. This is not as simple as the second quadratic term in s_q, but it has the benefit of not tending to dominate. S_2 is never larger than $2.5\sqrt{5}$, the value it has when $t = x$.

S_2 is needed specifically to penalise over-large values of s. With S_1 and S_3 alone, the rule is clearly not proper because the contestant can send S_1 to zero by making s extremely large. So S_1 encourages large values of s and it is the role of S_2 to penalise this in such a way as to make the rule proper. The coefficient $2.5\sqrt{5}$ is just the value needed to ensure the rule is proper (at least for a specific case, to be explored in more detail below).

Note that this shows that it is much simpler to balance the components in S than to balance the quadratic score. First, the balance between S_1 and S_2 has been set to achieve a proper rule (and, as we shall see, the value 2 in the exponent in S_2 has been chosen to ensure that the penalty for poor s is not too large). Second, we shall also see that to a large extent the balance between these two components and S_3 is not critical. As x moves far from t the score only grows quadratically, not as the fourth power.

The balance between questions is not difficult to achieve. S_1 and S_2 are automatically balanced by using the contestant's own s. An error by plus or minus one standard deviation, for example, gives the same value of these two components for any question. We need to choose $\bar{s}$ appropriately to each question, and a way of doing that will be described in the following development.

Finally, S is symmetric in $x - t$, so that an under-estimate gives the same score as an over-estimate by the same amount.

4 Propriety

We now consider whether S is indeed a proper scoring rule. Suppose that the contestant states values t' and s' as mean and standard deviation, while the contestant's true values of these are t and s. we wish to show that $E(S)$ is minimised by setting $t' = t$ and $s' = s$. We find the expectations of the first and third components of S easily.

$$\begin{aligned} E(S_1) &= E\left\{\frac{(X-t')^2}{s'^2}\right\} = (s')^{-2}\left\{(t'-t)^2 + s^2\right\} \ , \\ E(S_3) &= E\left\{\frac{(X-t')^2}{\bar{s}^2}\right\} = \bar{s}^{-2}\left\{(t'-t)^2 + s^2\right\} \ . \end{aligned}$$

However, we cannot evaluate the expectation of the third term so easily because it depends on the shape of the contestant's distribution of X, not simply on t and s.

4.1 Normal distribution

We begin by assuming that the distribution is normal, so that the contestant's actual distribution for X is $N(t, s^2)$. Then we can evaluate $E(S_3)$ and it is

helpful to take a more general case. Let $S_2 = a\exp\{-b(x-t')^2/s'^2\}$, where for the moment a and b are arbitrary positive constants. Then some standard manipulations give

$$\begin{aligned} E(S_2) &= E\left[a\exp\{-b(x-t')^2/s'^2\}\right] \\ &= \frac{as'}{\sqrt{s'^2+2bs^2}}\exp\left\{-b\frac{(t'-t)^2}{s'^2+2bs^2}\right\} . \end{aligned}$$

Hence

$$E(S) = \left\{(s')^{-2}+\bar{s}^{-2}\right\}\left\{(t'-t)^2+s^2\right\} + \frac{as'}{\sqrt{s'^2+2bs^2}}\exp\left\{-b\frac{(t'-t)^2}{s'^2+2bs^2}\right\} .$$

To minimise this with respect to t' and s', we differentiate with respect to both variables:

$$\frac{\partial E(S)}{\partial t'} = 2\left\{(s')^{-2}+\bar{s}^{-2}\right\}(t'-t) - 2\frac{abs'(t'-t)}{(s'^2+2bs^2)^{3/2}}\exp\left\{-b\frac{(t'-t)^2}{s'^2+2bs^2}\right\} ,$$

$$\begin{aligned} \frac{\partial E(S)}{\partial s'^2} &= -(s')^{-4}\left\{(t'-t)^2+s^2\right\} + \frac{abs^2}{s'(s'^2+2bs^2)^{3/2}}\exp\left\{-b\frac{(t'-t)^2}{s'^2+2bs^2}\right\} \\ &\quad +4\frac{ab^2s'(t'-t)^2}{(s'^2+2bs^2)^{5/2}}\exp\left\{-b\frac{(t'-t)^2}{s'^2+2bs^2}\right\} . \end{aligned}$$

The first clearly equals zero at $t'=t$, and the second is zero at $t'=t$ and $s'=s$ if and only if

$$(1+2b)^{3/2} = ab .$$

So we now impose the constraint that $a = (1+2b)^{3/2}/b$. In our actual scoring rule, $a = 2.5\sqrt{5}$ and $b=2$, which satisfies this constraint (and of course this is where the coefficient $2.5\sqrt{5}$ comes from).

However, this solution will only be a minimum if the matrix of second derivatives is positive definite. The second derivatives are readily found and evaluating them at $t'=t$ and $s'=s$ (and imposing the condition that $(1+2b)^{3/2}=ab$) gives the matrix

$$\begin{pmatrix} 2\bar{s}^{-2} & 0 \\ 0 & s^{-4}\{1+(b-1)/(1+2b)\} \end{pmatrix} ,$$

which is clearly positive definite when $b>0$, and in particular in our case where $b=2$.

This have proved that there is a minimum at $(t'=t, s'=s)$, where the minimum expected score is found to be

$$E(S)_{\min} = 3 + b^{-1} + s^2/\bar{s}^2 .$$

However, it is possible that this is not *the* minimum. In particular, we can see the role of the third component (which has played essentially no part in the above proof, except to ensure that the second derivative with respect to t' is positive) if we consider letting both t' and s' become large together so that $(x-t')^2/s'^2 \to 1$. Then we get $E(S_1+S_2) \to 1+b^{-1}(1+2b)^{3/2}e^{-b}$, and this is always less than $3+b^{-1}$. The third component prevents the contestant getting a better expected score by making both t' and s' large. Further exploration has indicated that $(t'=t, s'=s)$ is indeed the minimum for $E(S)$ provided $\bar{s} \leq 2s$.

4.2 Other distributions

We have not shown that S is a proper scoring rule for all possible distributions, just for the normal distribution. However, the expectation of S_2, which is the part which depends on the shape of the contestant's distribution for X, should in general be similar to the formula above. S_2 is symmetric around $x=t'$, and this should mean that the expectation is relatively unaffected if the distribution of X is asymmetric. Also S_2 decays rapidly towards zero, and so is insensitive to the tail probabilities.

It is possible that if, for instance, a contestant were to formulate beliefs about X in, say, a gamma distribution and numerically evaluated $E(S)$ over a range of t' and s' that some small improvement could be made on the expected score by setting $t' \neq t$ and/or $s' \neq s$. However, this hardly seems a realistic approach and we remain confident that the proposed scoring rule is suitable for its purpose.

4.3 Choice of b

The choice of $b=2$ was made again with a view to balance between the first score component (which is primarily concerned with t') and the second (which penalises large s'). At $b=2$, the range of possible values of S_2 is from 0 to $2.5\sqrt{5} = 5.59$, while the expected value is 2.5. This compares with values of S_1, which range from 0 to infinity, with an expected value of 1. Larger values of b seem to give S_2 too little influence because it too rapidly decays to zero, while smaller values give S_2 too much weight. This choice will be reviewed in the light of the results of the first Crystal Ball competition, and may change if the competition is run again in future.

5 Setting $\bar{s}$

The final unresolved question is how to set $\bar{s}$. Its value should be the same for all contestants, but can (and should) vary from one question to another. As indicated above, $\bar{s}$ should not be too large relative to a contestant's s, and indeed small values give more emphasis to getting the estimate t close to the actual x, which must be the main focus of the competition.

Our proposal is to set it, for each question, equal to the lower quartile of s values given to that question in the sample of contestants. This is another choice that will be reviewed in the light of experience in the first Crystal Ball competition.

6 Summary

The proposed scoring rule has been carefully chosen with a view to (a) being fair (i.e. proper), to the extent that it is possible to show this, (b) not giving too much emphasis to the assessment of s, and (c) being able to adjust weights to give all questions more or less equal value in the total score.

The scoring rule may not be unconditionally proper, but there seems to be little scope for distorting choices of t and s to get better expected scores. Anyway, the nature of the competition is that contestants win not be just getting their best possible scores on average but by beating other contestants. So even if the scoring rule was unconditionally proper there would still be some limited scope for improving winning chances by riskier choices of t and s.

Within the spirit of a competition which is supposed to be fun for contestants and observers alike, such nuances seem to be unimportant. We believe that the chosen scoring rule will do a good job of rewarding those who honestly and accurately assess their expectations and standard deviations for the uncertain quantities.