Rewarding Honesty vs Rewarding Accuracy: Computations with PIS and CRPS

Roger Cooke April 1, 2022, revised[1] June 12, 2022

Proper scoring rules were designed to encourage honesty in eliciting subjective probabilities. Rewarding honesty is not the same as rewarding accuracy. The simplest illustration of this is given by scoring *100* coin tosses for which an expert assesses the probability of Heads as ½. Using any of the familiar proper scoring rules, the score for the outcome Heads is the same as the score for the outcome Tails, and the score for *100* tosses is independent of the outcome sequence. *100* Heads gets the same score as *50* Heads, *50* Tails.

The Probability Interval Score *(PIS)* and its related Continuous Ranked Probability Scores (*CRPS)* have recently been applied to COVID-19 probabilistic predictions[2,3] and are discussed below. Numerical insight into these scores requires a bit of effort.

For the $(1-\alpha)$ interval $[L, U]$ with upper (lower) bound U (L), the *PIS* (negatively sensed) for realization y is $(U-L) + (2/\alpha)\times[(L-y)_+ + (y-U)_+]$ where $X_+ = X$ if $X > 0$ and $= 0$ otherwise. $s = 2/\alpha$ is the slope of the overconfidence penalty for $Y \notin [L,U]$. The length $(U\text{-}L)$ is called the "sharpness"; small values reward concentrated probability mass. If $Y\sim Unif[0,1]$, the central 0.*9* interval is $[0.05, 0.95]$ with expected *PIS*:

$$0.9 + 2 \times \int_{0..0.05} s\times u\, du = .9+(2/0.1)\times 0.05^2 = 0.95\ .$$

The integral is doubled to account for $Y > U$. The $1-\alpha$ interval need not be *central*; the interval $[0.1,1]$ is equally "sharp" and equally accurate statistically. However, the expected *PIS* is $0.9 + \int_{0..\,0.1} s\times u\, du$ (doubling the integration interval instead of the integral itself) = $0.9+20\times\ ½ \times 0.1^2= 1$. Although the sharpness and statistical accuracy are the same in these two examples, the expected interval scores differ. Suppose an expert prefers to give an *80%* interval $[0.1, 0.9]$, $s = 2/.2 = 10$. The expected score is $0.8 + 2\times\int_{0..0.1} s\times udu = 20 \times ½ \times 0.1^2 = 0.9 < 0.95$. An expert seeking to optimize (i.e.minimize) his/her expected score might take a central *2%* prediction interval $[0.49, 0.51]$ with expected score $0.02 +2\times2/.98\times0.49^2/2= 0.51$ (or take $lim_{\in\to 0}[,5 - \in,5 + \in]$ with expected score ½). All of these prediction intervals have zero information relative to the uniform background measure on $[0,1]$, so from that viewpoint there isn't much to choose.

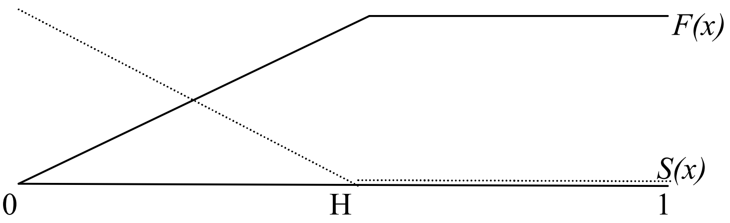
The way in which the *PIS* trades overconfidence for sharpness may strike some as counter-intuitive. For example, an expert claiming the degenerate interval [0.5, 0.5] has 40% probability of catching the realization would achieve an expected score of 0.833, better than the score of the 90% central interval. The sharpness of an interval of zero length outweighs the overconfidence of claiming 40% mass at the point 0.5. Of course this example is blocked if probability intervals are required to be 90%; assigning 90% mass to the point 0.5 returns an interval score of 5. Such scores from several experts could cause bad statistical performance, depending on how the experts are combined.

[1] This revision improves the exposition and combines discussions of *PIS* and *CRPS.*

[2] Ray, Evan L., et al, (2020), Ensemble Forecasts of Coronavirus Disease (COVID-19) in the U.S. medRxiv 2020.08.19.20177493; Posted August 22, 2020 doi: https://doi.org/10.1101/2020.08.19.20177493

[3] Cramer, Estee Y. Lopez, Velma K. +291 (2022), Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States PNAS April 8, 2022, https://doi.org/10.1073/pnas.2113561119

With several nested intervals the *PIS* converges to the *CRPS.* Consider realization Y ~ $Unif[0,1]$ and an assessment of Y by an expert whose distribution is $X \sim Unif[0,H],\ H \leq 1$. The expert thinks values $> H$ are impossible, although these can in fact arise. The expected *CRPS* is computed based not on what the expert believes, but on the realization Y. The *CDF* of X, $F(x) = x/H$ and the survivor function of X, $S(x) = 1\text{-}F(x)$ are pictured below:



We compute the expected Continuous Ranked Probability Score (*CRPS*) based not on what the expert believes, but on the realization Y. Consider an assessment of Y by an expert whose distribution is $X \sim Unif[0,H],\ H\leq 1$. The expert thinks values above H are impossible, although these can in fact arise.

The expected *CRPS* is $\int_{y\,=\,0..1}\ \int_{x\,=\,0...1} (F(x) - 1_{\{x\geq y\}})^2\,dx\,dy$. The calculation is broken into 2 steps:

$y \leq H$: $\int_{y=0..H}\ [\ \int_{x=0..y} (x/H)^2\,dx + \int_{x=y..H}((H-x)/H)^2\,dx\ ]\ dy$

$= \int_{y=0..H}\ [\ y^3/(3H^2)\ + \int_{z=H-y...0}\ z^2/H^2\ (-dz)\ ]\ dy\ =$

$= \int_{y=0..H}\ [\ y^3/(3H^2)\ + (H-y)^3/(3H^2)\ ]\ dy$

$=\ \ H^2/12\ + (1/(3H^2))\int_{z=H...0} z^3\ (-dz)\ dy\ = H^2/6.$

$y > H$: $\int_{y=H...1} [\ \int_{x=0..H} (x/H)^2 + \int_{x=H...y}\ dx + \int_{x=y...1} 0\ dx]\ dy$

$= \int_{y=H...1} [\ H/3 + y-H]\ dy$

$= H(1-H)/3 +\ \ \int_{0...1-H}\ z\ dz$

$= H(1-H)/3 + (1-H)^2/2.$

Therefore:

$E(CRPS(F,y)) = H^2/6 + H(1-H)/3 + (1-H)^2/2.$

If $X \sim Unif[L, H]\ ,\ 0 \leq L \leq H \leq 1,$ then the same method of calculation applies mutatis mutandis. If $L = 1-H$ then the contributions from $x \leq y$ and $y \leq x$ are equal and we need only double the contribution from x≤y. In that case, for $y \leq L,$ the contribution from $x \leq y$ is zero, since $x > L$. Therefore we compute

$\int_{y=\,L..H}\ \int_{x=L..y}\ \ F(x)^2 dx\ dy\ \ + \int_{y=H..1} \int_{x=L..y} F(x)^2 dx\ dy$

$$= \int_{y= L..H} \int_{x=L..y} (x-L)^2/(H-L)^2 dx\, dy + \int_{y=H..1} [\int_{x=L..H} (x-L)^2/(H-L)^2 dx + \int_{x=H..y} dx] dy$$

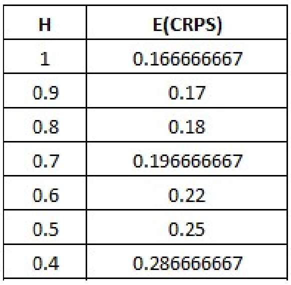
$$= \int_{y= L..H} (y^3/3)/(H-L)^2 dx + \int_{y=H..1} [(H-L)/3 + (y-H)]\, dy$$

$$= (H-L)^2/12 + (H-L)(1-H)/3 + (1-H)^2/2.$$

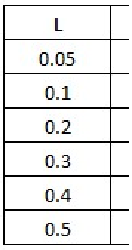
Adding the identical contribution from $y \leq x$ *gives:*

$$E(CRPS(F,y)) = (H-L)^2/6 + 2(H-L)(1-H)/3 + (1-H)^2.$$

Some values are



H	E(CRPS)
1	0.166666667
0.9	0.17
0.8	0.18
0.7	0.196666667
0.6	0.22
0.5	0.25
0.4	0.286666667



L	
0.05	
0.1	
0.2	
0.3	
0.4	
0.5	

Note that the expected CRPS for $X \sim uniform\ [0,H],\ H \geq 0.5$ is the same as that for $X \sim uniform\ [1 - H, H]$. Thus, $E(CRPS)$ for $X \sim uniform[0, 0.7] = 0.196\underline{6}\ = E(CRPS)$ for $X' \sim uniform[0.3, 0.7]$. An expert who believes X finding that 30% of the realizations Y are impossible has the same expected *CRPS* as an expert who believes X' finding 60% of the realizations impossible. This illustrates how the *CRPS* compensates loss of statistical accuracy with a gain in “sharpness”.