Comment on Evaluating epidemic forecasts in an interval format

Roger M. Cooke, March 1, 2022
Resources for the Future, Dept Math TU Delft (ret)

The authors are applauded for tackling challenging real problems. I especially welcome bringing performance data to bear on the expert judgment problem, as this raises the scientific potential of expert judgment. The following remarks on probability interval scoring should not obscure the agreement on these most important issues.

Scoring rules were introduced by de Finetti in 1937 as tools for encouraging honesty in eliciting subjective probabilities (1) and have been further developed by many authors (2). Experts receive a score as a function of their probability assessment and the realization. The score is strictly proper if experts maximize their expected score per item by, and only by, stating their true beliefs. Using a result of Murphy (1977) (3), DeGroot and Fienberg (1983) gave an additive decomposition of strictly proper rules into 'calibration' and 'refinement' terms (4), thereby replacing Murphy's 'resolution' (refinement applies only to well-calibrated experts).

Scoring rules for individual variables were not designed for evaluating or combining experts and are not generally fit for that purpose. Consider standard proper rules applied to 100-coin tosses. An expert assesses the probability of heads on each toss as ½. The score for the outcome heads is the same as the score for tails on each toss. If the score for all 100 assessments is a function of their 100 scores for the individual tosses, then their score for 100 tosses is independent of the outcome sequence; the outcome of 100 heads receives the same score as 50 heads and 50 tails. There are many counter intuitive examples (5, 6).

The authors apply scoring rules for probability intervals to COVID-19 probabilistic predictions (7). For the $(1-\alpha)$ interval $[L, U]$ with upper (lower) bound U (L), the (negatively sensed) score 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 $Y \not\in [L,U]$. The length $(U-L)$ 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 interval score:

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

The integral is doubled to account for $Y > U$. The $1-\alpha$ interval need not be centered; the interval $[0.1,1]$ is equally "sharp" and equally accurate statistically. However, the expected score is $0.9 + \int_{0..\ 0.1} s\times u\, du$ (doubling the integration interval instead of the integral itself) = 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 $[0.1, 0.9]$, $s = 2/.2$ and the expected score is 0.9. An expert seeking to optimize (i.e.minimize) his/her expected score might opt for a central 2% prediction interval $[0.49, 0.51]$ with expected score $0.02 + 2\times 2/.98\times 0.49^2/2 = 0.51$ (or take $\lim_{\in\to o}[0.5 - \in, 0.5 + \in]$ with expected score $\frac{1}{2}$). 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 interval score 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 expected interval score of *5*. Such scores from several experts could induce poor statistical performance, depending on how the experts are combined.

Scoring rules for average probabilities or equivalently, for expected relative frequencies, were designed to counter the issues raised here, for a recent exposition, see (8).

1. de Finetti, B. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* **7**, 1–68 (1937).
2. E. H. Shuford, A. Albert, H. Edward Massengill, Admissible probability measurement procedures. *Psychometrika* **31**, 125–145 (1966).
3. Murphy, A. H. The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review* **105**, 803–816 (1977).
4. DeGroot, M. H. and Fienberg S. E. The comparison and evaluation of forecasters. *The Statistician* **32**, 14–22 (1983).
5. Cooke, Roger M., (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science* Oxford University Press.
6. Cooke, Roger M. (2014) "Validating Expert Judgments with the Classical Model" in Experts and Consensus in Social Science - Critical Perspectives from Economics, Sociology, Politics, and Philosophy. Editors: Carlo Martini and Marcel Boumans, Series title: Ethical Economy - Studies in Economic Ethics and Philosophy, Springer
7. Johannes Bracher, Evan L. Ray, Tilmann Gneiting, Nicholas G. Reich (2021) Evaluating epidemic forecasts in an interval format, PLOS Computational Biology, Published: February 12, 2021 https://doi.org/10.1371/journal.pcbi.1008618/
8. Cooke, R. M., Marti, D., Mazzuchi, T., "Expert forecasting with and without uncertainty quantification and weighting: What do the data say?" *International Journal of Forecasting* **37**, 378–387 (2021).