Advances in Reliability, Safety and Security

ESREL 2024
Collection of Extended Abstracts

Scoring Rules And Performance Evaluated With Expert Judgment Data

Gabriela F. Nane[a], Roger M. Cooke[a,b]

[a]*Department of Applied Mathematics, Delft University of Technology, The Netherlands*
[b] *Resources for the Future, Washington, D.C.*

Keywords: expert judgment, scoring rules, Brier score, logarithmic score, continuous ranked probability score, probability interval score, mean absolute percentage error, geometric probability, Classical Model, overconfidence, location bias

A review of scoring rules highlights the distinction between rewarding honesty and rewarding quality. This motivates the introduction of a scale invariant version of the Continuous Ranked Probability Score (CRPS) which enables statistical accuracy testing based on an exact rather than an asymptotic distribution of the density of convolutions. A recent data set of 6,761 expert probabilistic forecasts for questions from their fields with realizations is used to compare performance. New insights include
Bulleted lists may be included and should look like this:

- variance due to assessed variables dominates variance due to experts;
- performance on Mean Absolute Percentage Error (MAPE) is weakly related to statistical accuracy;
- scale invariant CRPS combinations compete with the Classical Model on statistical accuracy and MAPE;
- CRPS unlike CM fails to detect location bias.

Introduction

Scoring rules were introduced by DeFinetti to reward honesty in expert elicitation. (De Finetti, 1937). Continuous Ranked Probability Scores (CRPS) (Brown, 1974), Probability Interval Scores (PIS) (Aitchison and Dunsmore, 1968) and scores from the Classical Model (CM) (Cooke, 1991) have recently captured attention in evaluating COVID-19 probabilistic model predictions (Ray et al., 2020; Cramer et al., 2022; Colonna et al., 2022), fueling the debate on how probabilistic predictions should be evaluated. This article offers insights from a recent expert judgment dataset (Cooke et al., 2021) comprised of expert probabilistic predictions over a wide variety of fields for which realizations or true values are also available.

Rewarding honesty in expert elicitation is not the same as rewarding quality in expert probabilistic assessments. This is manifested when numerically equal scores mask large differences in quality (Nane and Cooke 2024). In traditional proper scoring rules, statistical accuracy and some measure of informativeness (sharpness, resolution, refinement, information) are hard wired such that very high sharpness can buy off an attendant very poor statistical accuracy. In the Classical Model (CM) these are measured separately and combined in a product form with statistical accuracy strongly dominating. A scale invariant version of the CRPS isolates the statistical accuracy component and can be combined with informativeness as in the CM. This is applied to an expert judgment data base involving 49 studies, 526 experts and 580 calibration variables from their fields. With a closed form convolution of independent CRPS scores, the transformed CRPS yields a score for individual variables together with a test for experts' statistical accuracy on sets of variables without recourse to an asymptotic distribution. This may enable applications with fewer calibration variables. Compared to the statistical accuracy test used in the Classical Model it has the advantage of better rewarding proximity of a median point forecast to the realization. On the other hand, it is insensitive to location and under-confidence bias. Scoring

individual variables might prove useful for screening calibration variables for outliers. Tables 1 and 2 compare results for statistical accuracy and mean absolute percentage error.

Table 1. Statistical accuracy for different combinations (Decision Makers, DMs); this is the probability of falsely rejecting the hypothesis that the DM is statistically accurate. Low scores near 0 mean that's it is unlikely that the DM's probabilistic statements are true.

	EWDM	GWDM	$GWDM_{opt}$	IWDM	$IWDM_{opt}$	CRPS	Best MAPE expert
5%	0.04	0.02	0.02	0.01	0.01	0.01	0.00
50%	0.29	0.39	0.55	0.49	0.64	0.34	0.02
95%	0.65	0.66	0.93	0.83	0.96	0.65	0.70
mean	0.31	0.37	0.50	0.43	0.54	0.35	0.15
geomean	0.18	0.23	0.30	0.24	0.32	0.21	0.00

Table 2. MAPE scores for DMs giving the percentage error of medians as point forecasts; low scores near 0 are good.

	EWDM	GWDM	$GWDM_{opt}$	IWDM	$IWDM_{opt}$	CRPS	Best MAPE expert
5%	0.25	0.19	0.20	0.22	0.18	0.22	0.21
50%	0.84	0.65	0.74	0.55	0.60	0.65	0.54
95%	6.23	6.23	6.47	5.73	6.47	6.33	1.93
mean	3.64	2.33	2.50	1.64	1.90	2.05	0.94
geomean	0.95	0.90	0.95	0.73	0.84	0.82	0.58

References

Aitchison, J., Dunsmore, I. .1968. Linear-loss interval estimation of location and scale parameters, Biometrica 55(1),141–148.

Brown, T. A. 1974. Admissible scoring systems for continuous distributions. RAND Corporation.

Colonna, K. J., Nane, G. F., Choma, E., Cooke, R. M., Evans, J. S. 2022. A retrospective assessment of covid-19 model performance in the us. Royal Society Open Science.

Cooke, R. M. 1991. Experts in uncertainty: opinion and subjective probability in science. Oxford University.

Cooke, R. M., Marti, D., Mazzuchi., T. 2021. Expert forecasting with and without uncertainty quantification and weighting: What do the data say? International Journal of Forecasting 37(1), 378–387.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y. et al. 2022. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. Proceedings of the National Academy of Sciences 119(15), e2113561119.

De Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. In: Annales de l'institut Henri Poincaré 7, 1–68.

Nane, G., Cooke, R.M. 2024, Scoring Rules and Performance, New Analysis of Expert Judgment Data, appearing in Futures and Foresight Science.

Ray, E. L., Wattanachit, N., Niemi, J., Kanji, A. H., House, K., Cramer, E. Y., Bracher, J., Zheng, A., Yamana, T. K., Xiong, X. et al. 2022. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. MedRXiv, pages 2020–08, 2020.