FOCUS forecast scoring method

JHU/APL July 25, 2019

Overview

This memo describes how counterfactual forecasts will be scored in the IARPA FOCUS program. As stated in the test plan, the program goal for each performer is to exceed the comparison group performance with an effect size greater than or equal to a Cohen's d of 0.5. A Cohen's d measure requires a method of assigning scores, and a meaningful measure of standard deviation. Performer forecasts will be scored using a Ranked Probability Scoring method for ordered forecasts and a Brier score for unordered forecasts. A weighting scheme will be used to create equivalent expected scores among forecasts with different numbers of response bins. A square root transformation will be applied to these weighted scores to improve psychometric properties. A pooled standard deviation can then be used to calculated effect sizes, as described in the power analysis.

Problem framing

Scoring forecasts in FOCUS has presented some unexpected challenges due to fact that (a) both forecasts and ground truth take the form of probability distributions; and (b) these forecasts are elicited across response formats that vary according to their number of bins. Comparing a distribution with a distribution tends to lower squared difference scores overall, creating a highly skewed distribution of scores. The differing number of response bins leads to different variances between problems, potentially leading to problem sets with different effects for each problem. These problems are not unique to FOCUS. However, most previous geopolitical forecasting markets have evaluated forecasts against single outcomes (i.e. observed values in the real world) rather than distributions of outcomes from a set of ground truth worlds. Scoring a forecast distribution against an outcome distribution accentuates the problem of skewed scores. FOCUS also used a greater variety of response types to accommodate different shapes and types of responses.

As these issues became apparent through pre-testing and comparison group testing, the government and T&E team began to examine various scoring alternatives. Since performer teams also included researchers with considerable experience and insight in this area, a scoring working group was formed including representatives from IARPA, the JHU/APL T&E team, and each of the three performers teams. A number of alternatives were considered over the course of several months. A preliminary scoring plan was produced and used in FOCUS cycle 1. The scoring group continued to examine alternatives and provide input. The method described in this memo is identical to that used in Cycle 1 with the addition of a square root transformation of the weighted forecasts, which was added for reasons that will be described.

Scoring of ordered forecasts

Performer forecasts on questions with ordered responses will be scored using Ranked Probability Scoring (RPS). Ranked probability scores can be used to score ordered multinomial outcomes where subjects get credit for being “closer” to the true outcome. As discussed in

Constantinou & Fenton (2012), the ranked probability score is both *strictly proper* and *sensitive to distance*, making it an appropriate method for scoring the accuracy of ordinal probability forecasts (see also: Epstein, 1969; Murphy, 1969; Murphy, 1970). In general, the ranked probability score is defined as:

$$RPS = \frac{1}{r-1}\sum_{i=1}^{r}\left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j\right)^2$$

Where *r* is the number of potential outcomes, and p_j and e_j are the forecasts and observed outcomes at position *j*. A worked example of RPS was given at kickoff in a slide deck on scoring (slide 9). The RPS is very similar to a method recommended by Jose, Nau and Winkler (2009), which has been used in prior IARPA forecasting programs for ordered forecasting problems. The only differences between these two approaches are the mathematical procedures by which they are calculated (their results are perfectly correlated) and the fact that the RPS is scaled from 0-1 instead of 0-2.

Scoring of unordered forecasts

Unordered forecasts will be scored with a mean squared error. This could be considered an average Brier score for multiple bin answers. These will be referred to simply as a Brier scores. Binary responses can be scored either way with identical results.

Weighting based on number of response bins

RPS and Brier scores derived from responses with larger numbers of responses bins tend to have lower means and standard deviations than responses from fewer bins (e.g. binary responses.) This creates a psychometric problem, in that some problems tend to influence the mean score more than others.

A different method of addressing this problem, which was considered but is not being used, is to use scores standardized for each question (i.e. z-scores). However, in this dataset, the skewed distributions and small number of data points available for normalization made this less viable.

FOCUS will instead use a weighting method to create problems with roughly equivalent expected scores. Each RPS score and Brier score will be multiplied by a coefficient to adjust for the different means and variances of scores on items with different numbers of bins. The coefficients were derived from simulated data where random uniform ground truth and forecasts (summing to 100) were scored 10 million times across different numbers of bins.

# of Bins	Average RPS Score	Weight-Multiplier
2	0.083	1
3	0.056	0.083/0.056=1.482
4	0.042	1.98
5	0.033	2.52
6	0.028	2.96
7	0.024	3.46

8	0.021	3.95
9	0.019	4.37
10	0.017	4.89

Table 1. Weightings for ordered forecasts by number of bins

# of Bins	Average Brier	Weight-Multiplier
2	0.083	1
3	0.055	1.52
4	0.037	2.22
5	0.027	3.13
6	0.02	4.20
7	0.016	5.46
8	0.012	6.88
9	0.0099	8.44
10	0.0082	10.2

Table 2. Weightings for unordered forecasts by number of bins

Square root transformation of scores

Weighted RPS and Brier scores will be transformed with a square root function. Figures 1 and 2 show the effects of the square root transformation on the overall distribution of Cycle 1 scores. The square root transformation has several desirable properties. First, it reduces the overall skew of the distribution of scores, bringing it closer to a normal distribution. Being closer to normal makes the standard deviation, upon which effect size measurements are based, more meaningful and interpretable. Second, the square root transformation makes the distribution more symmetrical. In a highly skewed distribution, extreme values on one side (in this case, poor forecasts on the far right) can have a disproportionate effect on mean scores. On the left, score differences become smaller and smaller as they approach the minimum of zero. The danger is that substantive improvements in forecast quality might make very little difference to the overall mean score, and thus not be rewarded by the scoring method.

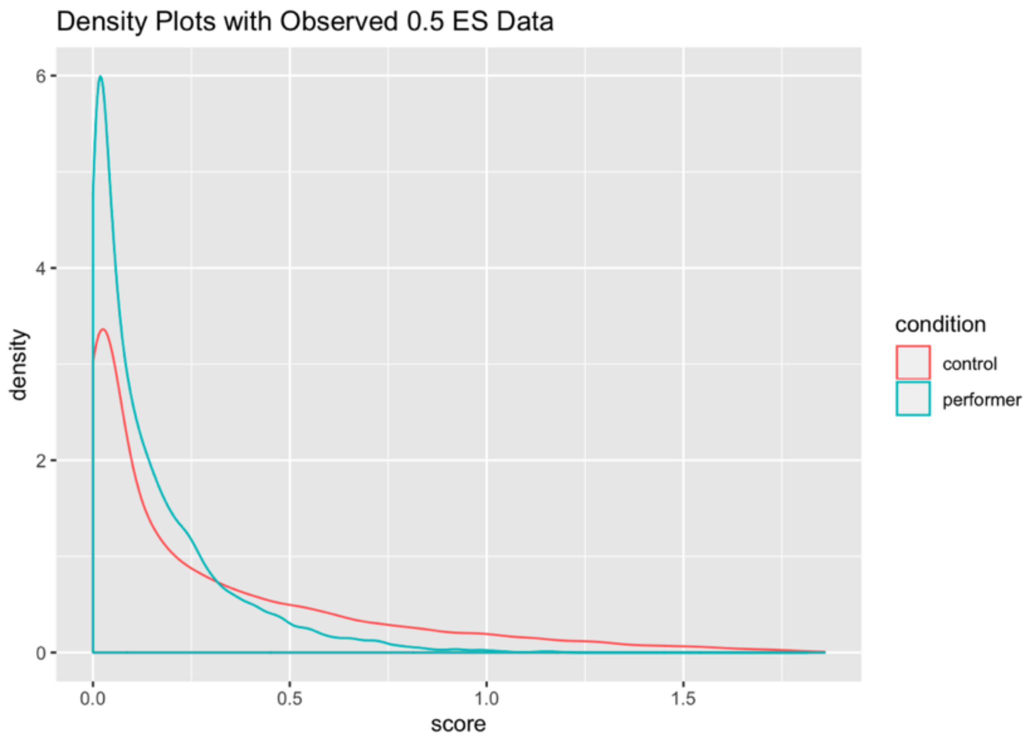

Figure 1. Simulated forecasts from comparison group and successful performer

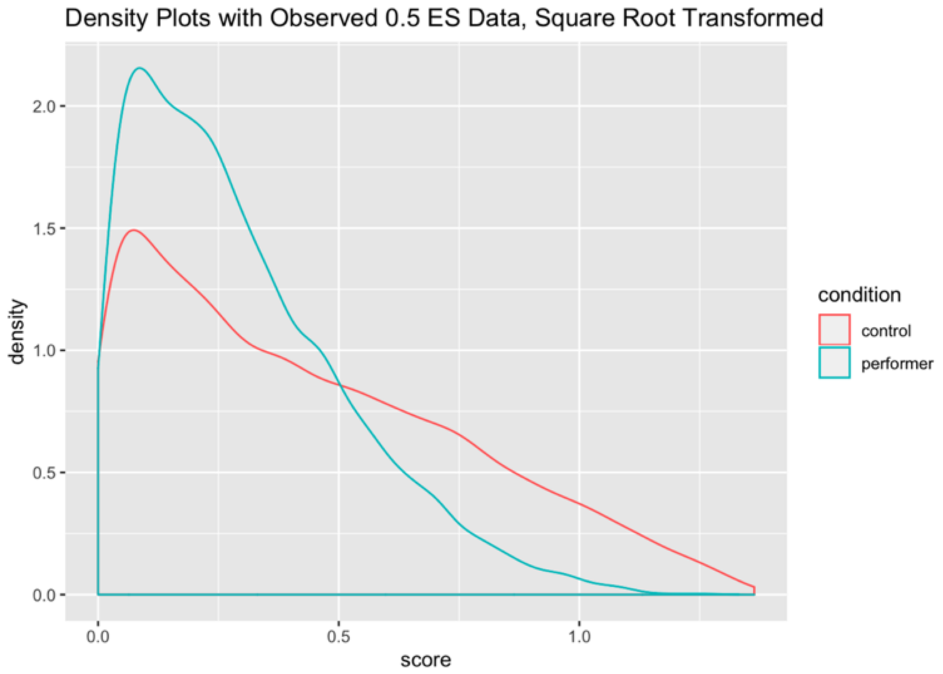

Figure 2. Square root transformed data of simulated forecasts from comparison group and successful performer

The scoring working group considered other alternative transformations to address this, including normalization and two nonparametric methods, including using Kendall's Tau for scoring individual forecasts by comparing bin order, and a version of Kruskal-Wallis for comparing RPS scores. The Kruskal-Wallis, which converted RPS scores into problem rankings (e.g. each problem's 13 forecasts were given ranks 1-13) went even further than the square root transformation in reducing effects of outliers on each side. One topic of discussion was whether in real-world intelligence applications very large 'misses' in forecasting should be weighted more highly than incremental improvements on average forecasts. This topic warrants further study from mathematical, psychometric, and analytic tradecraft perspectives.

The square root transformation was judged to be a good compromise, improving psychometric properties while preserving some of this asymmetry. The square root transformation also has some face validity, in that its overall effect is to put forecasts back on the same scale as the original distance scores (distance between ground truth and forecast in each bin) before these differences were squared and summed.

A lingering issue related to square root transformation was whether its use made scores no longer strictly proper. This is likely to be the case, theoretically opening the possibility that some sort of strategy which does not incentive the reporting of true beliefs might be used to improve scores. However, simulation exercises performed by the T&E team did not find a clear example of such a strategy. This remains an open question.

Effect size calculations

The scores obtained by the methods described above will then be compared to the distribution of scores obtained from the FOCUS comparison groups. Cohen's d will be computed using pooled standard deviations of comparison group and performer data.

Alternate effect size calculations were considered, include Hedge's g and deriving Cohen's d from a transformation of Kendall's Tau.

Future developments

It is the intention of the T&E team that this scoring method will be used for the duration of the FOCUS program, including future cycles and any replication testing. This should be considered the official metric for judging whether the effect size target has been reached. As described in the test plan, however, success in FOCUS will not be defined by any single measure, but by considering all relevant and available quantitative and qualitative data.

The government may request and be provided additional analyses by T&E, which may include different scoring methods, alternative data transformations, different weighting schemes, different effect size measures, etc. Performers may also suggest additional measures or provide additional analyses for the government's consideration which may be relevant to understand the strengths and weaknesses of different forecasting approaches.

References

Constantinou, A. C., & Fenton, N. E. (2012). Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models. *Journal of Quantitative Analysis in Sports*, 8(1). https://doi.org/10.1515/1559-0410.1418

Epstein, E.S. (1969). A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology*, 8. 985-987.

Hedges, L.V. (1981). "Distribution theory for Glass' estimator of effect size and related estimators". Journal of Educational Statistics. 6 (2): 107–128. doi:10.3102/10769986006002107.

Kendall, Maurice; Gibbons, Jean Dickinson (1990) [First published 1948]. Rank Correlation Methods. Charles Griffin Book Series (5th ed.). Oxford: Oxford University Press. ISBN 978-0195208375.

Murphy, A. H. (1971). A note on the ranked probability score. *Journal of Applied Meteorology* 10(1) 155–156.

Murphy, A. H. (1973), A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595-600

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. Journal of Educational and behavioral Statistics, 23(2), 170-192.