MITRE

Hybrid Forecasting Competition (HFC)

Brier Score Calculations

Authors:
Dr. Rob Hartman
Phil Hilliard
Jon Whitenack

August 2020

This publication is based upon work supported by the Intelligence Advanced Research Projects Activity (IARPA) [IARPA-BAA-16-02], via contract 2015-14120200002-002, and is subject to the Rights in Data-General Clause 52.227-14, Alt. IV (May 2014). Any views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation therein.

©2020 The MITRE Corporation. All rights reserved.

Approved for Public Release; Distribution Unlimited. Public Release Case Number 20-2080

McLean, VA

Table of Contents

1 Brier Score Calculations 1-1
1.1 Core Scoring Rule 1-1
1.2 Properness of Mean Daily Brier (MDB) Scores 1-1
1.3 Application of Brier Scoring to Ordered Categorical Forecast Outcomes 1-2
2 References 2-1

1 Brier Score Calculations

1.1 Core Scoring Rule

The Hybrid Forecasting Competition used the Brier score (aka Mean Quadratic Score or MQS) as our primary forecast accuracy scoring rule.

This document describes how to calculate the Brier score from the forecasts in the prediction sets and daily forecasts files.

This proper scoring rule is applicable to probabilistic forecasts offered in discrete choice scenarios (i.e., How likely is each specified outcome?). It can be calculated as:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}(fij - oij)^2$$

where

- f_{ij} = the forecast probability for outcome option j and o_{ij} = 1 (if outcome option j comes to pass) or 0 (otherwise).
- r = the number of possible forecast outcomes
- n = number of forecasts.

Under Brier's (1950) original formulation, these scores can range from 0 to 2, with lower scores indicating superior accuracy (i.e., minimal deviation from observed outcomes).

Readers will note that for binary forecasting problems this score is equivalent to a widely-used version of the Brier score that ranges from zero to one (0-1), but unlike that version, Brier's (1950) original formula is also applicable to unordered multinomial probabilistic forecasts.

1.2 Properness of Mean Daily Brier (MDB) Scores

The Brier score is generally a proper scoring rule, meaning that the optimal strategy for obtaining the best score is to report one's honest beliefs. However, for Individual Forecasting Problems (IFPs) whose resolution may occur prior to some deadline (e.g., Will Party A attack Party Y before the end of the month?), the mean of a set of daily Brier scores (the mean daily Brier (MDB)) is no longer proper: systems are incentivized to exaggerate the probability of the focal event occurring (vs. not occurring). Various possible solutions exist to restore properness, but all of those identified to date have significant drawbacks, including diminished power to detect differences between systems or complications in forecast elicitation or interpretation. For these reasons, HFC retained the MDB metric in HFC. Despite the lack of properness for this subset of IFPs (expected to be a minority), the use of MDBs gives no forecasting system an inherent advantage over any other, such that differences in relative accuracy should be preserved.

1.3 Application of Brier Scoring to Ordered Categorical Forecast Outcomes

Some forecasting questions require the assignment of probabilities to ordered categories. Consider, for example, a question whose outcome possibilities are temporal intervals:

a. 20 days or less
b. Between 21 and 40 days
c. Between 41 and 60 days
d. 61 days or more.

We call these ordered individual forecasting problems, or oIFPs.

In this example, if the event occurs during period B, the following distribution

D1: A – 25%, B – 25%, C – 50%, D – 0%

is in principle more accurate than

D2: A – 25%, B – 25%, C – 30%, D – 20%

because D1 is uniformly closer to the truth than D2, even though both assigned p=0.25 to the correct interval.

Jose, Nau and Winkler (2009) identify extensions of binary and unordered multinomial scoring rules to ordered categorical forecasts. Their approach satisfies the following criteria.

1. If a distribution is uniformly closer to ground truth than a second distribution, then the first distribution receives a better score.
2. The scoring rule is strictly proper.
3. The scoring rule is commensurate with the Brier score in terms of scaling and interpretation.

The basic procedure is as follows.

Step 1 – Take the original categories (A-B-C-D) and break them up into a set of cumulative, binary categories (A-BCD; AB-CD; ABC-D)

Step 2 – Apply the scoring rule to each binary pair

Step 3 – Take an average across the binary pair scores.

In the above example, the standard unordered Brier scores for D1 and D2 are 0.88 and 0.76, respectively[1]. Note that even though D1 was closer to ground truth than D2 it received a worse

[1] For D1, Brier= ((1-0.25)^2) + ((0-0.25)^2) + ((0-0.50)^2) + ((0-0.00)^2) = 0.875. For D2, Brier= ((1-0.25)^2) + ((0-0.25)^2) + ((0-0.30)^2) + ((0-0.20)^2) = 0.755.

score. This occurred because the Brier score increases as the distribution over the false categories deviate from uniform. In general, we would expect that any forecast system that assigns a high probability to a near miss will be penalized using the original Brier score.

In contrast to the unordered results, the 'ordered' Brier scores for D1 and D2 are 0.21 and 0.24 respectively[2]. The closer distribution receives a better score, which is what we would desire in this scenario.

Finally, two implications of this approach should be noted. First, any credit given to near misses may lead to a circumstance where a higher probability assigned to the ground truth category may not lead to the highest score. For example, consider the following distributions where B is ground truth

D1: A – 0%, B – 50%, C – 25%, D – 25%

D2: A – 20%, B – 30%, C – 40%, D – 10%

The original Brier score for D1 and D2 are 0.38 and 0.70. D1 scores considerably better (lower). The revised Brier scores are 0.21 and 0.20. D2 receives a better (lower) score even though it assigns 20% less to the ground truth category. This reflects the ordered scoring rule's focus on the overall extent to which one's probability mass function is tightly centered around the category that contains the true quantity, which can be contrasted with the traditional multinomial Brier score's (questionable in this case) property that any probability not assigned to the true outcome should be uniformly (i.e. indifferently) distributed among the remaining (incorrect) outcome options.

[2] For D1, Brier for A-BCD = ((1-0.75)^2)+((0-0.25)^2)=0.125, Brier for AB-CD=((1-0.50)^2)+((0-0.50)^2)=0.5, Brier for ABC-D=((1-1.00)^2)+((0-0)^2)=0.00. Ordered Brier is the average of these: 0.208. For D2, Brier for A-BCD = ((1-0.75)^2)+((0-0.25)^2)=0.125, Brier for AB-CD=((1-0.50)^2)+((0-0.50)^2)=0.5, Brier for ABC-D=((1-0.80)^2)+((0-0.20)^2)=0.08. Ordered Brier is the average of these: 0.235.

2 References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78(1),* 1-3.

Jose V., Nau R., Winkler R. (2009), Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science, 55(4),* 582-590