A Hypothesis Test for Brier Scores

Elliot Christophers

Abstract

Bradley et al. (2008) produced an expression for the variance in Brier scores (Brier, 1950), which required both the probabilistic forecasts and the observed outcomes. We provide a complementary set of equations, where the mean and variance of Brier scores can be derived using only the probabilities. By demonstrating that our expressions are accurate, we present a framework where a forecaster can readily evaluate whether their model is calibrated or whether it should be reconsidered.

1 Introduction

The Brier score (Brier, 1950) is a strictly proper scoring rule for quantifying the strength of a vector of probabilistic forecasts, given their outcomes. Bradley et al. (2008) developed – and Wilks (2010) discussed further – an expression for the variance of Brier scores, incorporating both the forecasts and their outcomes. These two authors demonstrate the accuracy of the expression, illustrating with Monte Carlo simulation that the randomly sampled variances match the predicted variances across a broad range of forecasts. Although useful, for instance, in defining a confidence interval for the forecast-generating model's Brier skill score, what the Bradley et al. framework fails to accomplish is testing whether the model is calibrated or not, in the absence of some target or benchmark Brier score. We supplement this deficiency by providing estimates for the Brier score mean and variance using only the probabilistic forecasts, allowing us to use the observed Brier score as a test statistic to compare against this estimated distribution. In our paper, we provide these estimates for the mean and variance of the true Brier score under the null hypothesis of perfect calibration, and proceed to evaluate the accuracy and suitability of these expressions using Monte Carlo simulation.

2 Brier Score Hypothesis Testing Framework

We have an n-tuple of probability forecasts $\boldsymbol{f} = (f_1, \ldots, f_n)$ for n events, and the n-tuple of outcomes for these events $\boldsymbol{x} = (x_1, \ldots, x_n)$, where each x_i is a Bernoulli trial with unknown parameter, $x_i \in \{0, 1\}\ \forall\ i$. We can calculate $\hat{S}$, an estimate for the Brier score S (Brier, 1950) of these forecasts, using

$$\hat{S} = \frac{1}{n}\sum_{i=1}^{n}(f_i - x_i)^2 \tag{1}$$

where the Brier score is a strictly proper scoring rule (Gneiting and Raftery, 2007), incentivising the forecaster to provide their true belief f_i for each forecast i. The Brier score is a loss function, ranging from 0 to 1 with 0 representing perfect forecasting.

We want to develop a hypothesis test for whether our model, from which $\boldsymbol{f}$ is generated, is calibrated. We define this null hypothesis as

$$H_0 : E[x_i] = f_i\ \forall\ i \tag{2}$$

that is, the expected value of the i-th outcome variable x_i equals the model forecast: $x_i \sim \text{Bernoulli}(f_i)$. In this framework, we assume that $\boldsymbol{f}$ is calibrated, and hence calculate the distribution of sample Brier scores that would be observed under the null hypothesis. We calculate

our observed Brier score using (1) by combining $\boldsymbol{f}$ with $\boldsymbol{x}$. We then compare this test statistic to the null distribution, and evaluate (2) at a given significance level.

The expected Brier score $E[S]$ is calculated

$$E[S] = E\left[\frac{1}{n}\sum_{i=1}^{n}(f_i - x_i)^2\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[(f_i - x_i)^2\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[f_i^2 - 2f_i x_i + x_i^2\right] = \\ \frac{1}{n}\sum_{i=1}^{n}\left(E\left[f_i^2\right] + E\left[-2f_i x_i\right] + E\left[x_i^2\right]\right) = \frac{1}{n}\sum_{i=1}^{n}(f_i^2 - 2f_i^2 + f_i) = \frac{1}{n}\sum_{i=1}^{n}(f_i - f_i^2) \tag{3}$$

using that, under the null, the f_i are constant, and that, with the x_i being Bernoulli random variables, $x_i^2 = x_i$.

We can similarly calculate the variance of the Brier score, $Var[S]$, as

$$Var[S] = E[S^2] - E[S]^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n}(f_i - x_i)^2\right)^2\right] - \left(\frac{1}{n}\sum_{i=1}^{n}(f_i - f_i^2)\right)^2 \tag{4}$$

Expanding, $\left(\frac{1}{n}\sum_{i=1}^{n}(f_i - x_i)^2\right)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n}(f_i - x_i)^2\right)\left(\sum_{j=1}^{n}(f_j - x_j)^2\right) =$
$\frac{1}{n^2}\left(\sum_{i=1}^{n}(f_i - x_i)^4 + \sum_{i\neq j}(f_i - x_i)^2(f_j - x_j)^2\right)$. Hence, $E[S^2] =$
$\frac{1}{n^2}\left(\sum_{i=1}^{n} E\left[(f_i - x_i)^4\right] + \sum_{i\neq j} E\left[(f_i - x_i)^2(f_j - x_j)^2\right]\right) =$
$\frac{1}{n^2}\left(\sum_{i=1}^{n} E\left[(f_i - x_i)^4\right] + \sum_{i\neq j} E\left[(f_i - x_i)^2\right] E\left[(f_j - x_j)^2\right]\right)$. We have that $E\left[(f_i - x_i)^4\right] =$
$f_i(1-f_i)^4 + (1-f_i)f_i^4$ and that $E\left[(f_i - x_i)^2\right] = f_i - f_i^2$, which gives $E[S^2] =$
$\frac{1}{n^2}\left(\sum_{i=1}^{n}\left[f_i(1-f_i)^4 + (1-f_i)f_i^4\right] + \sum_{i\neq j}\left[(f_i - f_i^2)(f_j - f_j^2)\right]\right)$, and finally

$$Var[S] = \frac{1}{n^2}\left(\sum_{i=1}^{n}\left[f_i(1-f_i)^4 + (1-f_i)f_i^4\right] + \sum_{i\neq j}\left[(f_i - f_i^2)(f_j - f_j^2)\right]\right) - \left(\frac{1}{n}\sum_{i=1}^{n}(f_i - f_i^2)\right)^2 \tag{5}$$

Thus, given $\boldsymbol{f}$, we can use (3) and (5) to calculate the mean and variance of the distribution of Brier scores, under the null hypothesis in (2). As desired, these expressions are independent of $\boldsymbol{x}$, which only enters into the calculation of our test statistic from (1). With the Brier score bounded by 0 and 1, it seems reasonable to assume that these sample Brier scores are beta distributed $S \sim \text{Beta}(v, w)$ (Bayes, 1763; Krzysztofowicz and Long, 1991), with probability density function

$$\frac{S^{v-1}(1-S)^{w-1}}{B(v,w)} \tag{6}$$

where B is the beta function and v and w are parameters such that if a beta distribution has mean μ and variance σ^2, then

$$v = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right) \\ w = \frac{1-\mu}{\mu}v \tag{7}$$

Thus, with $\mu = E[S]$ from (3) and $\sigma^2 = Var[S]$ from (5), using (7) we can calculate the v_S and w_S which govern the distribution of Brier scores under the null hypothesis in (2). For our test statistic $\hat{S}$ from (1) using $\boldsymbol{f}$ and $\boldsymbol{x}$, we obtain a p-value

$$p_S = 1 - I_{\hat{S}}(v_S, w_S) \tag{8}$$

where $I_{\hat{S}}$ is the regularised incomplete beta function, which is the cumulative distribution function of the beta distribution, evaluated at $\hat{S}$ (Peizer and Pratt, 1968). Since the Brier score is a strictly proper scoring rule, our hypothesis test only considers the positive tail, as this is where the scores from uncalibrated models would appear: the negative tail consists of calibrated models which have experienced fortunate sequences of outcomes, with respect to their forecasts.

The next section considers the accuracy of our expressions in (3) and (5), the validity of the assumption that the Brier scores are beta distributed, and the performance of (8) as a solution to (2).

3 Analysis

To evaluate the accuracy of our equations for $E[S]$ and $Var[S]$, as well as the beta assumption, we run a Monte Carlo simulation. In each of $m = 10,000$ iterations, we draw $n = 10^u$ where $u \sim U(\log_{10}(50), 3)$ – that is, n is drawn on a logarithmic scale between 50 and 1,000. We also draw parameters for a beta distribution $v_f, w_f \sim U(0.5, 5)$, giving $\boldsymbol{f} = (f_1, \ldots, f_n \mid f_i \sim \text{Beta}(v_f, w_f)\ \forall\ i)$, from which we calculate $E[S]$ and $Var[S]$ using (3) and (5), respectively. The bounds for the v_f and w_f are chosen to give a relatively broad range of distributions of forecasts, to mimic the breadth of possible forecasting scenarios which may occur in practice. Then, with $\boldsymbol{f}$, we generate 10,000 n-length vectors $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj} \mid x_{ij} \sim \text{Bernoulli}(f_i)\ \forall\ i)$, and for each of these $\boldsymbol{x}_j$ calculate the corresponding Brier score $\hat{S}_j$ using (1). Thus, for each iteration's $\boldsymbol{f}$, we have 10,000 Brier-scores, which follow the distribution described under the null hypothesis in (2).

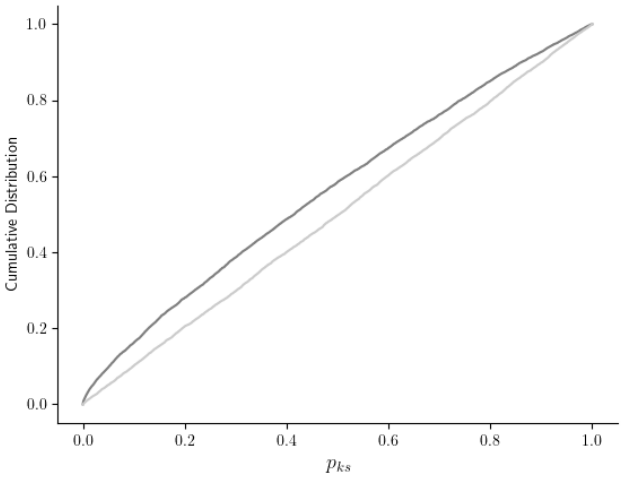

Figure 1: The cdf of Kolmogorov-Smirnov p-values across 10,000 iterations. The darker line corresponds to our expressions for $E[S]$ and $Var[S]$, while the lighter line tests data drawn from the specified distribution.

We want to evaluate, first, whether $E[S] \approx \bar{S}$ and $Var[S] \approx s_S^2$, that is, whether our expressions for the mean and variance match the sample values across the $\hat{S}_j$; and second, whether the $\hat{S}_j \sim \text{Beta}(v_S, w_S)$, where v_S and w_S are calculated from (7) using $E[S]$ and $Var[S]$ (note, not using $\bar{S}$ and s_S^2: we want to ascertain that the $\hat{S}_j$ are beta distributed with our expressions for the mean and variance, not merely with the observed ones). To accomplish the former evaluation, we simply calculate Pearson's correlation coefficient (Pearson, 1895) across the 10,000 $(E[S], \bar{S})$ and $(Var[S], s_S^2)$ pairs, respectively. For the latter, we use the one-sample Kolmogorov-Smirnov (KS) test (An, 1933; Smirnov, 1948), a non-parametric test which gives the probability of a sample being drawn from a certain continuous probability distribution with specified parameterisation, with the alternative hypothesis being that the data is drawn from some other distribution.

We find that $\text{Corr}(E[S], \bar{S}) = 0.999993$ and $\text{Corr}(Var[S], s_S^2) = 0.999753$, with a practi-

cally perfect fit between the predicted values from our equations (3) and (5). Indeed, when these are plotted, the resulting figures are barely distinguishable from a straight line through the origin with unit slope. The results from the KS tests are also convincing, with the null hypothesis being sustained in 90.3% and 96.8% of the $m = 10,000$ iterations, at 5% and 1% significance levels, respectively. The darker line in Figure 1 shows the cumulative distribution of the p-values p_{ks} from the 10,000 iterations. The lighter line is the cumulative distribution of p-values from a KS test when the $\hat{S}_j$ values are drawn immediately from the beta distributions with v_S and w_S, and gives the cumulative uniform distribution which p-values follow under the null, as expected. Hence, the concavity of the dark line demonstrates that there is a greater proportion of smaller p-values than one would expect if the $\hat{S}_j$ were indeed beta distributed with parameters v_S, w_s, suggesting that our expressions for the mean and variance do not give perfect fits. However, for our purposes, testing the hypothesis in (2), it should be sufficient. Indeed, Figure 2 plots nine sample distributions of 10,000 $\hat{S}_j$ as well as the beta distribution (dark line) parameterised by v_S, w_S, for cases where $p_{ks} < 0.001$, which occurred in only 0.68% of iterations. The visual fit to the general shape of the distribution is perfect, with the disparity arising from arbitrary spikes at certain values of $\hat{S}_j$. Given our use case, where we will compare an observed Brier score $\hat{S}$ to its distribution under the null hypothesis in (2) according to (8), this approximate fit is more than adequate.

We now evaluate whether the KS p-value depends on the distribution of the $\boldsymbol{f}$ or its number of elements n. As described above, in each iteration we randomly draw n probabilities $f_i \sim \text{Beta}(v_f, w_f)$. We use beta regression (Kieschnick and McCullough, 2003; Ferrari and Cribari-Neto, 2004; Geissinger et al., 2022) to regress the p_{ks} on linear combinations of n, v_f, w_f and their higher polynomials. Essentially, beta regression takes a conventional multi-linear regression model and transforms it using, in this case, a logit link function, whereby the model predictions are bounded by 0 and 1. We can thus evaluate the extent to which n, v_f and w_f contribute to p_{ks}. We use the Bayesian information criterion (Schwarz, 1978), given by

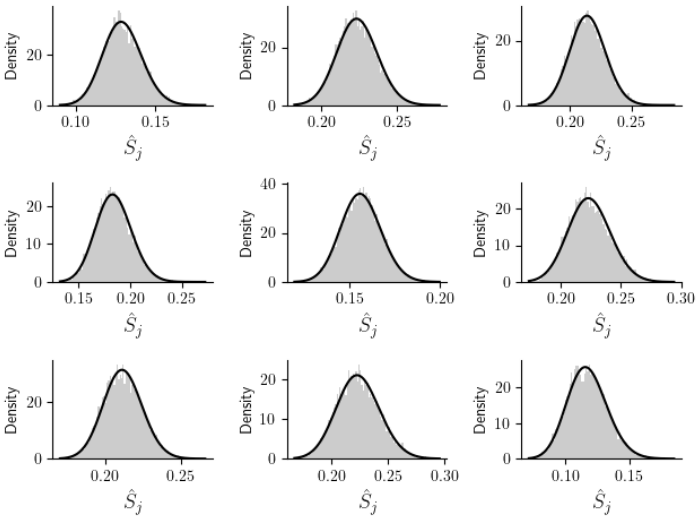

Figure 2: Nine sample cases of predicted (dark) and realised distributions of $\hat{S}_j$ where the Kolmogorov-Smirnov p-value is $p_{ks} < 0.001$. Even for these poorly-fit cases, the visual match is very good.

$$\text{BIC} = k \ln(m) - 2 \ln(L) \tag{9}$$

where k is the number of model parameters, m is the data sample size, and L is the maximised likelihood of the model, upon optimisation. The BIC provides a nuanced comparison of different models, penalising unnecessary parameters (unduly high k) while rewarding improved model performance (higher L). Relatively, the more negative is BIC, the better is the model. For a base model with only a constant, we find that BIC $= -724.93$. A second model with a constant as well as the linear and quadratic terms of v_f and w_f has BIC $= -753.76$, suggesting that these two parameters contribute valuable information in attempting to predict p_{ks}. Adding the linear

and quadratic terms of n to this model, however, gives BIC $= -1105.42$, a vast outperformance, demonstrating that the majority of variation in p_{ks} is explained by n, the number of forecasts made.

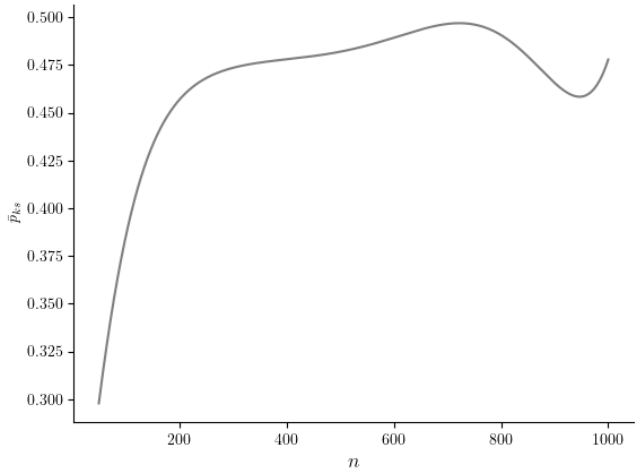

Figure 3: Using Beta regression, we model the Kolmogorov-Smirnov p-values p_{ks} as a function of n and its higher-order polynomial terms.

Figure 3 plots p_{ks} as a function merely of n and its polynomials up to degree seven. Obviously this model is overfit, but our intention is to explain the observed data to the utmost extent, not to make a-priori predictions. We observe that, above $n \approx 200$, the p-values are practically constant, with a mean of 0.475 or so, while there is a sharp increase from $n = 50$, suggesting that the majority of cases where our expressions for $E[S]$ and $Var[S]$ did not fit the observed data occurred when $n < 200$. Of course, even at $n = 50$, the mean value $\bar{p}_{ks}$ is roughly 0.3, so on average our calculations are valid, and indeed, Figure 2 suggests that even when the fit is bad, per se, it is still more than serviceable for our purposes.

We hence conclude that for two n-tuples of forecasts $\boldsymbol{f}$ and realised outcomes $\boldsymbol{x}$, the observed Brier score $\hat{S}$ from (1) follows a beta distribution with mean given by (3) and variance given by (5) under the null hypothesis in (2) of the forecast-generating model being perfectly calibrated. We can then use (8) to evaluate whether the model is indeed calibrated, at a chosen significance level. To demonstrate the test, we perform a second Monte Carlo simulation. In each of 1,000 iterations, we draw $v_f, w_f \sim U(0.5, 5)$. Then, in iteration, we draw n as above 10,000 times and for each generate $\boldsymbol{f} = (f_1, \ldots, f_n \mid f_i \sim \text{Beta}(v_f, w_f) \ \forall \ i)$. For each $\boldsymbol{f}$, we calculate $E[S]$ and $Var[S]$. To generate $\boldsymbol{x}$ we use the linear model of Murphy and Wilks (1998), adopted by Bradley et al. (2008), given by

$$E[x_i] = (1 - \Delta) f_i + \Delta \frac{v_f}{v_f + w_f} \tag{10}$$

where $\Delta \geq 0$ is a parameter and $\frac{v_f}{v_f + w_f}$ is the mean of the beta distribution of the f_i. If $\Delta = 0$, then (10) gives $E[x_i]$ as (2) would demand, while if $\Delta > 0$, then the forecasts are less individually calibrated, increasingly becoming the base-rate frequency, the climatological probability. For different values of Δ, we generate $\boldsymbol{x}$ corresponding to $\boldsymbol{f}$ according to (10). For different values of the significance level α, we then evaluate the null hypothesis in (2) using (8). Each of the 1,000 iterations will then have, for each (α, Δ)-pair, 10,000 observations of whether the null hypothesis in (2) is rejected or not. For each iteration, we take the proportion of cases in which the null is sustained. Table 1 gives $1 - barp_{\alpha\Delta}$ the mean of these proportions across the iterations, for each (α, Δ)-pair.

Considering first the case where $\Delta = 0$, we would expect to see a proportion α of cases leading to rejections, since we are under the null hypothesis. With a slight bias, this is indeed what we observe. The test seems to be somewhat over-sensitive, rejecting a proportion slightly

greater than α. But clearly it is very well-calibrated overall. Indeed, as Δ increases, the proportion or rejections increases at each significance level, as expected, since model-calibration decreases in Δ.

This relationship is actually extremely predictable: if we denote $1-\bar{p}_{\alpha\Delta}$ as the mean proportion of rejections across the 1,000 iterations, with the given values of Δ and α, if we regress $\bar{p}_{\alpha\Delta} = \beta_0+\beta_1\Delta+\beta_2\alpha+\varepsilon_{\alpha\Delta}$ by using ordinary least squares, we get an R-squared metric of 0.956. This, of course, is merely across the nine observations shown in Table 1, but if we instead index the 1,000 iterations by i and then denote $1-p_{i\alpha\Delta}$ as the proportion of rejections for pair $(\alpha\Delta)$ in the i-th iteration and regress $p_{i\alpha\Delta} = \beta_0 + \beta_1\Delta_i + \beta_2\alpha_i + \varepsilon_{i\alpha\Delta}$, we instead have 9,000 observations and obtain an R-squared of 0.775: not quite as perfect, but representative of the strong relationship between the three variables nonetheless, especially given that there are only nine pairs of values to regress on.

		α		
		0.01	0.05	0.1
	0	0.989	0.949	0.899
Δ	0.125	0.898	0.759	0.652
	0.25	0.695	0.512	0.407

Table 1: $1-\bar{p}_{\alpha\Delta}$: the mean proportion of cases where the null hypothesis in (2) is sustained using (8), as a function of Δ and α.

We conclude this section by stating, then, that our hypothesis test developed above is robust and calibrated across a large range of values for the number of forecasts n as well as the distribution of these forecasts $\boldsymbol{f}$. A forecaster can thus, after making a moderate number of predictions and observing the outcomes, evaluate whether their model is calibrated or not.

4 Test Eligibility

We have seen above that the beta fit is good even for $n = 50$, but this was for distributions of probabilities f_i ranging from $v_f, w_f \in [0.5, 5]$. In this section we suggest a rule for determining whether the test is applicable. The criteria for a normal approximation to the beta distribution being accurate relate, in general, to the approximate equality of the parameters v and w, or to their size (Wise, 1960). Here, however, we will imagine that a related criterion is applicable, namely, that $E[S] - z\sqrt{Var[S]} \geq 0$, which comes from the rule for a normal approximation being appropriate for the binomial distribution, where $z = 3$. We can rearrange to find $\frac{E[S]}{\sqrt{Var[S]}} \geq z$.

Figure 4 plots the usefulness of this rule at different values of z, for a Monte Carlo simulation very similar

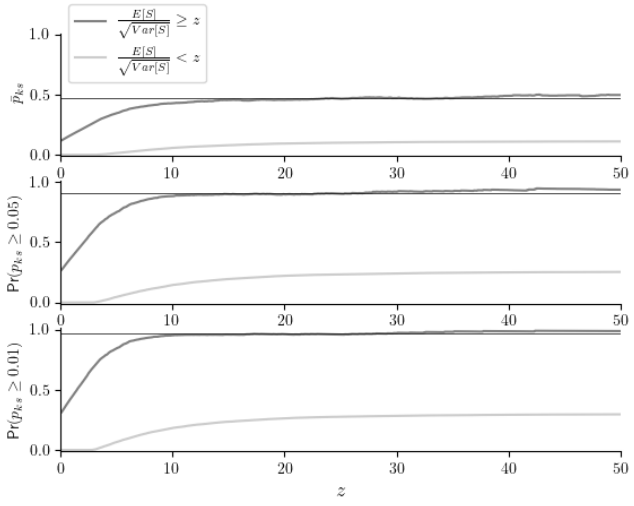

Figure 4: Filtering by $\frac{E[S]}{\sqrt{Var[S]}}$ (dark lines aggregate above z), displaying the mean p-value (upper panel) and the probability of sustaining the KS null hypothesis. The horizontal lines denote usual behaviour, attained at $z \approx 10$.

to that above, the only difference being that we draw n on a logarithmic scale between 5 and 1,000, and v_f and w_f on a logarithmic scale between 0.1 and 100, to give a broader range of possible forecasting scenarios. In the three panels, the darker line considers those iterations which gave $\frac{E[S]}{\sqrt{Var[S]}} \geq z$, while the lighter line shows the complement. The upper panel gives the mean KS p-value for these two groups. The second panel gives the proportion of p-values above and below 0.05, while the lower panel does the same but for the significance level 0.01. The horizontal lines in each are drawn at 0.475, 0.903 and 0.968. These numbers were mentioned above as the representative values when the mean and variance estimates were good and the beta distribution behaved well. We can see that this normalcy is attained roughly for $z = 10$, demonstrating the ability of this rule to filter out the cases where the test is not quite applicable.

We want to extract n from this relation, giving a threshold sample size. We can express the numerator as $E[S] = E[f_i - f_i^2]$. Separating the individual components of $Var[S]$, we find $\frac{1}{n^2}\sum_{i=1}^{n}\left[f_i(1-f_i)^4 + (1-f_i)f_i^4\right] = \frac{1}{n}E\left[f_i(1-f_i)^4 + (1-f_i)f_i^4\right]$ and $\frac{1}{n^2}\sum_{i\neq j}\left[(f_i - f_i^2)(f_j - f_j^2)\right] \approx \frac{n-1}{n}\left(E[f_i - f_i^2]\right)^2$ giving $Var[S] \approx \frac{1}{n}E\left[f_i(1-f_i)^4 + (1-f_i)f_i^4\right] + \frac{n-1}{n}\left(E[f_i - f_i^2]\right)^2 - \left(E[f_i - f_i^2]\right)^2 = \frac{1}{n}\left(E\left[f_i(1-f_i)^4 + (1-f_i)f_i^4\right] - \left(E[f_i - f_i^2]\right)^2\right)$. Thus,

$$\frac{E[S]}{\sqrt{Var[S]}} \approx \frac{E[f_i - f_i^2]}{\sqrt{\frac{1}{n}\left(E\left[f_i(1-f_i)^4 + (1-f_i)f_i^4\right] - \left(E[f_i - f_i^2]\right)^2\right)}} \tag{11}$$

If we assume that the distribution of $\boldsymbol{f}$ remains constant, this implies that $\frac{E[S]}{\sqrt{Var[S]}} \underset{\approx}{\propto} \sqrt{n}$. We thus obtain the heuristic $n_\tau = z^2\frac{Var[S]}{E[S]^2}$, where n_τ is the threshold value of n which allows us to apply the hypothesis test. However, when this rule was used instead of that in Figure 4, the output was surprisingly dissimilar, despite the correlation between the left and right hand sides in (11) being 0.9998 in the simulation. Indeed, for the cases where $\frac{E[S]}{\sqrt{Var[S]}} < 10$, 54.5% had $n \geq n_\tau$. We suggest, then, that the applicability of our test has more to do with the distribution of the forecast probabilities and the ensuing distribution of Brier scores than it does the sample size, and that if our rule for eligibility is not fulfilled, simply gathering more forecasts will not ensure that it will be applicable eventually. This is not contradictory to the evidence in Figure 3, as we merely showed there that larger n improved the fit for cases which were already applicable.

5 Conclusion

In this paper, we have derived a framework complementary to that of Bradley et al. (2008), where highly accurate estimates of the mean Brier score and its variance can be calculated using only the vector of forecasts $\boldsymbol{f}$, where Bradley et al. required the outcome vector $\boldsymbol{x}$ as well. This development allows a forecaster to evaluate the hypothesis that their model is calibrated, in the sense that the expected value of the outcome variable x_i for the i-th forecast equals the probability f_i provided for this event, $E[x_i] = f_i$. After a reasonable number of predictions, the forecaster can calculate the observed Brier score $\hat{S}$ and compare it to the null distribution of Brier scores, possibly deciding, at a given significance level, that $\hat{S}$ is significantly above the mean $E[S]$, thus rejecting the null hypothesis that the model is calibrated, and indicating that the forecaster would be well-advised to return to the model design and consider why forecasts are thus inaccurate.

6 References

An, Kolmogorov. “Sulla determinazione empirica di una legge didistribuzione”. *Giorn Dell’inst Ital Degli Att*, vol. 4, 1933, pp. 89–91.

Bayes, Thomas. “LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S”. *Philosophical transactions of the Royal Society of London*, no. 53, 1763, pp. 370–418.

Bradley, A., et al. “Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score”. *Weather and Forecasting*, vol. 23, 2008, pp. 992–1006. https://doi.org/10.1175/2007WAF2007049.1.

Brier, Glenn W. “Verification of forecasts expressed in terms of probability”. *Monthly weather review*, vol. 78, no. 1, 1950, pp. 1–3.

Ferrari, Silvia, and Francisco Cribari-Neto. “Beta regression for modelling rates and proportions”. *Journal of applied statistics*, vol. 31, no. 7, 2004, pp. 799–815.

Geissinger, Emilie A, et al. “A case for beta regression in the natural sciences”. *Ecosphere*, vol. 13, no. 2, 2022, e3940.

Gneiting, Tilmann, and Adrian E Raftery. “Strictly proper scoring rules, prediction, and estimation”. *Journal of the American statistical Association*, vol. 102, no. 477, 2007, pp. 359–78.

Kieschnick, Robert, and Bruce D McCullough. “Regression analysis of variates observed on (0, 1): percentages, proportions and fractions”. *Statistical modelling*, vol. 3, no. 3, 2003, pp. 193–213.

Krzysztofowicz, Roman, and Dou Long. “Beta likelihood models of probabilistic forecasts”. *International Journal of Forecasting*, vol. 7, no. 1, 1991, pp. 47–55.

Murphy, Allan H, and Daniel S Wilks. “A case study of the use of statistical models in forecast verification: Precipitation probability forecasts”. *Weather and Forecasting*, vol. 13, no. 3, 1998, pp. 795–810.

Pearson, K. Notes on regression and inheritance in the case of two parents proceedings of the royal society of London, Vol. 58. 1895.

Peizer, David B, and John W Pratt. “A normal approximation for binomial, F, beta, and other common, related tail probabilities, I”. *Journal of the American Statistical Association*, vol. 63, no. 324, 1968, pp. 1416–56.

Schwarz, Gideon. “Estimating the dimension of a model”. *The annals of statistics*, 1978, pp. 461–64.

Smirnov, Nickolay. “Table for estimating the goodness of fit of empirical distributions”. *The annals of mathematical statistics*, vol. 19, no. 2, 1948, pp. 279–81.

Wilks, Daniel S. “Sampling distributions of the Brier score and Brier skill score under serial dependence”. *Quarterly Journal of the Royal Meteorological Society*, vol. 136, no. 653, 2010, pp. 2109–18.

Wise, ME. “On normalizing the incomplete beta-function for fitting to dose-response curves”. *Biometrika*, vol. 47, nos. 1–2, 1960, pp. 173–75.