International Journal of Research and Development in Engineering Sciences
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 8, Issue. 6 (November – December) , 2026 , Pages: 58 – 67
Received: 01 March 2026 ; *Accepted:* 30 October 2026 ; *Published:* 16 November 2026
https://doi.org/10.63328/IJRDES-V8RI6P6
https://www.youtube.com/watch?v=Z3g-1VyX_yI
https://archive.org/details/jsp-ijrdes-v8ri6p6
Original Paper : Deductive Research – Experimental Research Paper

Analytical Characterization of Finite-Sample Bias in Expected Calibration Error for Probabilistic Classifiers using Artificial Intelligence

Mesfin Dagne

Department of Department of Computer Science and Engineering , Adama Science and Technology University, Adama, Ethiopia

* *Corresponding Author : Mesfin Dagne;* *mesfindagne@gmail.com*

Abstract: The expected calibration error (ECE) is the standard number for reporting how well a classifier's probabilities match observed frequencies, and it is computed from a finite sample by binning. Because the estimate is an average of absolute values of noisy bin differences, it is biased upwards, and the bias is large at the sample sizes in which models are often evaluated. We derive the expected value of the binned estimate in closed form: each bin contributes the mean of a folded normal distribution, which gives $\Sigma w_b F(\delta_b, \sigma_b)$, and for a perfectly calibrated model the estimate has mean $s\sqrt{(2B/(\pi n))}$ with B bins and n cases. Over 75 simulated settings the formula was within 5.3% of the simulated mean (average 1.3%). A perfectly calibrated model evaluated on 1,000 cases with 15 bins has an expected ECE of 3.8%, and keeping it below 1 point needs about 9,900 cases with 10 bins. The sample size is proportional to the number of bins, so the choice of bins trades bias from noise against bias from binning, and the root mean squared error is smallest at a small number of bins (3, 5 and 12 for 300, 1,000 and 5,000 cases in our setting). Correcting the noise is not straightforward: subtracting the expected noise works for calibrated models but understates the error of miscalibrated ones by up to 3 points, while inverting the folded normal in each bin removes most of the bias for miscalibrated models, is not uniformly better in root mean squared error, and is worse for calibrated ones. On the breast cancer and digits datasets, the estimate for a model with weak features was 16.2% with 50 cases and 6.0% on all cases. Synthetic results are simulations; the real data are two small benchmark sets.

Keywords: Calibration, Expected Calibration Error, Binning, Bias, Folded Normal, Sample Size.

1. Introduction

A classifier that says "70% probability" should be right about 70% of the time. The expected calibration error (ECE) measures the gap: group predictions into bins by predicted probability, compare the average prediction in each bin with the observed frequency of the positive class, and average the absolute gaps [1, 2]. It is reported for neural networks [2], for classical models [3, 11], and in most work on recalibration [10].

The estimate is easy to compute, and its sampling behaviour is rarely examined. Each bin's observed frequency is noisy, the gap is an absolute value, and the absolute value of a noisy zero is positive. A model that is perfectly calibrated therefore has a positive expected ECE, and the amount depends on the sample size and the number of bins [5, 7, 17]. This is the same effect that makes the mean of $|N(0, \sigma^2)|$ positive, and it can be calculated. We calculate it. The expected binned ECE is the average over bins of the mean of a folded normal distribution, which has a closed form [12]. We then use the formula to answer practical questions: how large is the bias, how many cases are needed to make it small, how should the number of bins be chosen, and can the bias be removed? We make five contributions.

- A closed form for the expected value of the binned ECE, for a perfectly calibrated model and for a miscalibrated one, verified in simulation (Section 4).
- The sample size needed for a given bias, proportional to the number of bins and to the inverse square of the target.
- The trade-off in the number of bins between noise bias and binning bias, with the number of bins that minimizes the error.
- A comparison of four estimators, showing that simple noise corrections help for

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

calibrated models and hurt for miscalibrated ones.

- A real-data check on five models fitted to the breast cancer and digits datasets.

All synthetic experiments are simulations from known models. The real-data experiment uses cross-validated predictions on two small benchmark datasets.

2. Background and Related Work

Calibration and ECE. Calibration has long been studied in forecasting [4, 9, 8]. Naeini and colleagues defined the binned ECE for machine learning [1], and Guo and colleagues showed that modern neural networks are often overconfident and that temperature scaling repairs it [2]. Platt scaling and isotonic regression are older repairs [10, 3].

Estimation problems. Several authors note that the binned estimate depends on the number of bins and has a positive bias for finite samples [5, 6, 7, 17]. Kumar and colleagues gave an unbiased estimator for the squared calibration error and showed that the plug-in estimate is biased [5]. Nixon and colleagues compared binning schemes empirically [6], Vaicenavicius and colleagues proposed calibration tests [7], and Widmann and colleagues unified them [16]. Bröcker analysed the bias of binned reliability estimates in forecasting [17].

What we add. The existence of the bias is known. We contribute its exact expression in terms of folded normal, a sample-size rule, the optimal number of bins in a simulated setting, and a comparison of corrections for the absolute-value ($\ell 1$) error that is reported in practice.

3. Research Methodology and Gap

3.1. Calibration Error And Its Binned Estimate

A model assigns each case a probability q of the positive class, and the label is $y \in \{0, 1\}$. The model is calibrated if $\Pr(y = 1 \mid q) = q$. The calibration error is $\mathrm{ECE} = E\,|\Pr(y = 1 \mid q) - q|$. With B equal-mass bins, each holding $n_b = n/B$ cases, the estimate is

$$\mathrm{ECE_hat} = \Sigma_b\, (n_b/n)\, |\bar{y}_b - \bar{q}_b|. \qquad (1)$$

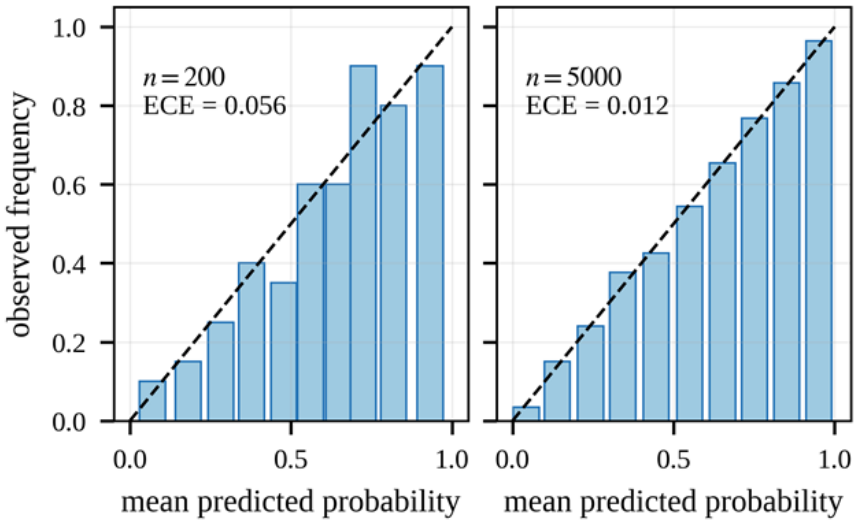


Figure. 1 Reliability diagrams of a perfectly calibrated model (uniform scores, 10 equal-mass bins). With 200 cases the bars scatter around the diagonal and the binned ECE is 0.056; with 5,000 cases it is 0.012. The true calibration error is zero in both.

3.2. Data and models

In the calibrated case the true probability p has one of three distributions, uniform on (0, 1), bimodal Beta(0.3, 0.3), and low-risk Beta(1, 6), the model reports $q = p$, and $y \sim \mathrm{Bernoulli}(p)$. For miscalibrated models the true log-odds are $z \sim N(0, 2^2)$, $p = \mathrm{logistic}(z)$, and the model reports $q = \mathrm{logistic}(a\ z)$; the model is overconfident for $a > 1$ and calibrated for $a = 1$. The true calibration error of these models is $E|q - p|$, which we compute from a sample of one million.

3.3. Four estimators

Plain: Eq. (1). **Null-corrected:** Eq. (1) minus the expected noise $\sqrt{(2/\pi)}\ \Sigma\ w_b\ \hat{\sigma_b}$, clipped at zero, where $\hat{\sigma_b}^2$ is the estimated variance of $\bar{y}_b$. **Soft threshold:** $\Sigma\ w_b\ \max(|\hat{d_b}| - \hat{\sigma_b}, 0)$, where $\hat{d_b} = \bar{y}_b - \bar{q}_b$. **Inversion:** $\Sigma\ w_b\ \tilde{\delta_b}$, where $\tilde{\delta_b}$ solves $F(\tilde{\delta_b}, \hat{\sigma_b}) = |\hat{d_b}|$ (zero if $|\hat{d_b}| \le \hat{\sigma_b}\sqrt{(2/\pi)}$).

4. Proposed System

Proposition 1 (expected ECE). Condition on the predicted probabilities. In bin b let δ_b be the mean of p − q over its cases and $\sigma_b^2 = \Sigma\ p(1 - p)/n_b^2$ its variance of $\bar{y}_b$, where $p = \Pr(y = 1 \mid q)$. If $\bar{y}_b - \bar{q}_b$ is approximately $N(\delta_b, \sigma_b^2)$, then

$$E[\mathrm{ECE_hat}] \approx \Sigma_b\ w_b\ F(\delta_b, \sigma_b), \quad F(\delta, \sigma) = \sigma\sqrt{(2/\pi)}\ e^{-\delta^2/(2\sigma^2)} + \delta[1 - 2\Phi(-\delta/\sigma)], \qquad (2)$$

where F is the mean of $|N(\delta, \sigma^2)|$, the folded normal [12], and $w_b = n_b/n$.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

Corollary 1 (calibrated model). *If the model is calibrated, $\delta_b = 0$ and $F = \sigma_b\sqrt{(2/\pi)}$. With equal-mass bins, $\sigma_b^2 = v_b B/n$ with v_b = mean of $p(1-p)$ in the bin, so*

$$E[\text{ECE_hat}] \approx \bar{s}\sqrt{(2B/(\pi n))}, \quad \bar{s} = (1/B)\,\Sigma_b \sqrt{v_b}. \tag{3}$$

For uniform scores $\bar{s} \approx 0.40$. The bias decreases as $1/\sqrt{n}$ and increases as $\sqrt{B}$. In general $F(\delta, \sigma) \geq |\delta|$ and $F(\delta, \sigma) - |\delta| \leq \sigma\sqrt{(2/\pi)}$, so the estimate never lies below the binned population error and never exceeds it by more than the calibrated-case bias.

Corollary 2 (sample size). *For a calibrated model the bias is at most ε when*

$$n \geq (2B/\pi)\,(\bar{s}/\varepsilon)^2. \tag{4}$$

The requirement is proportional to the number of bins and to $1/\varepsilon^2$: halving the target multiplies it by four.

Corollary 3 (equal-width bins). *With B equal-width bins of probability mass w_b and mean $p(1-p)$ equal to v_b, the same argument gives, for a calibrated model,*

$$E[\text{ECE_hat}] \approx \sqrt{(2/(\pi n))}\,\Sigma_b \sqrt{(w_b\, v_b)}. \tag{5}$$

The bias now depends on how the scores are spread over the bins: bins with large mass and large v_b dominate.

Remark (the number of bins). Equation (2) holds for each B. Increasing B reduces the binning error (the population error of a binned model is below the true error, because gaps of opposite sign inside a bin cancel) and increases the noise bias, which grows as √B. The total error of the estimate is smallest at an intermediate B that grows with n.

5. Experimental Setup and Implementation

For the calibrated case we use $n \in \{100, 300, 1{,}000, 3{,}000, 10{,}000\}$ and $B \in \{5, 10, 15, 20, 50\}$, with 600 repetitions per setting and three score distributions. For miscalibrated models we use $a \in \{1, 1.25, 1.5, 2\}$, $B = 15$, $n \in \{200, 1{,}000, 5{,}000\}$ and 600 repetitions. The required-n check searches a geometric grid of 14 sample sizes from 100 to 60,000 with 150 repetitions per point (60 for the largest), with $B \in \{10, 20\}$ and $\varepsilon \in \{0.02, 0.01\}$. The study of the number of bins uses $a = 1.5$, $B \in \{3, \ldots, 100\}$ and 300 repetitions. The estimator comparison uses $n = 1{,}000$, $B = 15$ and $n = 200$, $B = 10$ with 800 repetitions. The formula values use bin means and variances computed from a population of one million cases.

Real data are the Wisconsin breast cancer set (569 cases, 212 malignant) [14] and digits (1,797 cases, 180 are the digit 9) [15], from scikit-learn [13]. Predicted probabilities come from five-fold cross-validation for five models: logistic regression on all features (breast cancer and digits 9), logistic regression on two weak features (mean smoothness and mean symmetry), and Gaussian naive Bayes on all features (both datasets). Each subsample of $n \in \{50, 100, 200, 400\}$ cases is drawn without replacement, with 1,000 repetitions and $B = 10$. The calibrated-floor curve is computed from each subsample's own predicted probabilities as $\sqrt{(2/\pi)}$ times the mean of $\sqrt{(\text{mean } q(1-q)/n_b)}$, the value Eq. (3) would give if the model were calibrated.

6. Results and Analysis

6.1. A calibrated model has a positive ECE

Figure 1 shows the effect, and Table 1 and Figure 2 check Eq. (3). Over the 75 settings the formula is within 5.3% of the simulated mean, with average relative error 1.3%. A model that is perfectly calibrated has an expected ECE of 12.4% at 100 cases, 3.8% at 1,000 cases and 1.2% at 10,000, with 15 bins and uniform scores. The decrease follows the $1/\sqrt{n}$ ine in Figure 3. The bias depends on the score distribution through $\bar{s}$: it is 2.4% for bimodal scores, where most predictions are near 0 or 1 and have little variance, and 3.0% for low-risk scores. It increases with the number of bins (Figure 4), as $\sqrt{B}$.

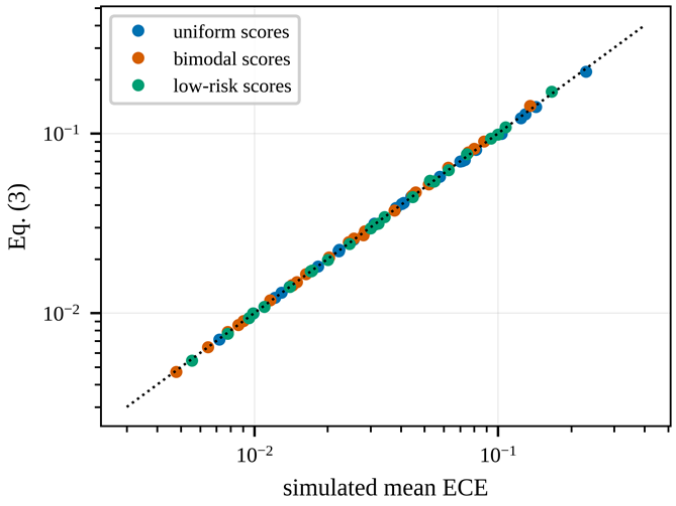


Figure. 2 Simulated mean ECE of a calibrated model against Eq. (3) for 75 settings (three score distributions, n from 100 to 10,000, B from 5 to 50). The dotted line is equality.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

Table. 1 Mean ECE (%) of a calibrated model ($B = 15$): simulation (600 repetitions) and Eq. (3).

Scores	n	Simulated	Eq. (3)	Rel. error (%)
uniform	100	12.4	12.2	2.1
uniform	1000	3.8	3.9	0.8
uniform	10000	1.2	1.2	0.3
bimodal	100	7.6	7.9	4.0
bimodal	1000	2.4	2.5	1.8
bimodal	10000	0.8	0.8	0.8
low-risk	100	9.4	9.4	0.2
low-risk	1000	3.0	3.0	1.1
low-risk	10000	1.0	0.9	1.3

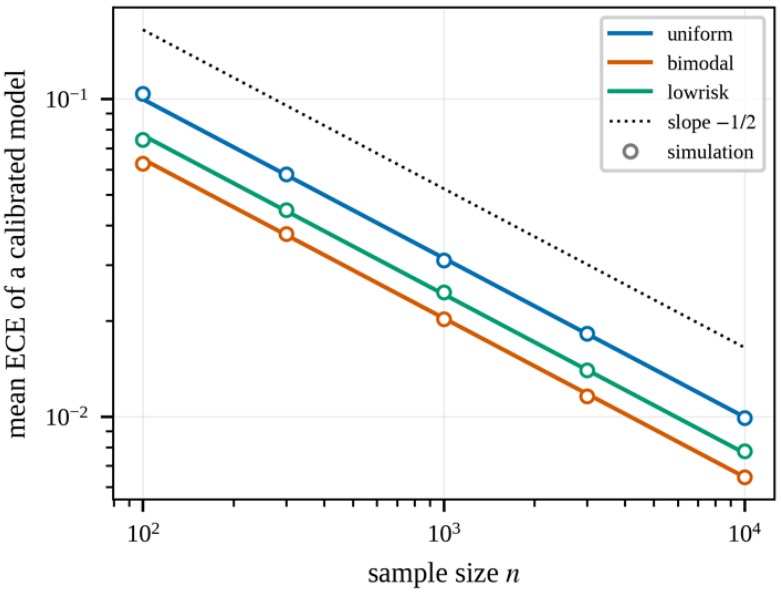


Figure. 3 Mean ECE of a calibrated model against sample size ($B = 10$). Lines: Eq. (3). Circles: simulation. The dotted line has slope −1/2.

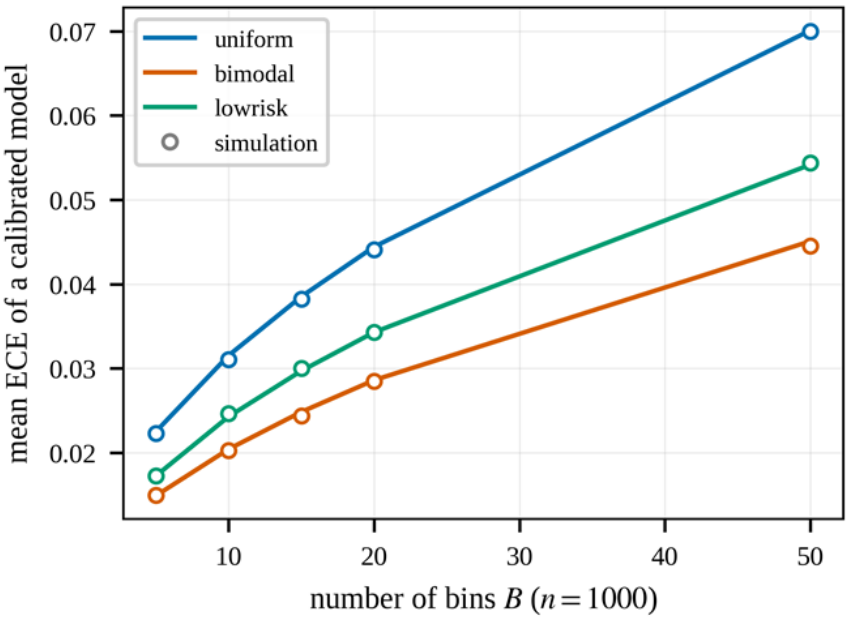


Figure. 4 Mean ECE of a calibrated model against the number of bins (n = 1,000). Lines: Eq. (3). Circles: simulation.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

6.2. Miscalibrated models

Table 2 and Figure 5 give models of increasing overconfidence. Eq. (2) is within 0.004 of the simulated mean in all 12 settings. At 200 cases the estimate is 8.4% for a model whose true calibration error is 3.3%, an overestimate by a factor of 2.6. At 1,000 cases it is 4.6%, and at 5,000 cases 3.5%, close to the true value. The bias shrinks as the true error grows (Corollary 1): for $a = 2$ the true error is 9.3% and the estimate at 1,000 cases is 9.6%. Models that are badly miscalibrated are measured well; models that are almost calibrated are not.

Table 2: Mean ECE estimate (%) of overconfident models ($B = 15$): true error, simulation and Eq. (2).

a	n	True	Simulated	Eq. (2)	Bias
1.00	200	0.0	8.3	8.1	8.3
1.00	1000	0.0	3.6	3.6	3.6
1.00	5000	0.0	1.6	1.6	1.6
1.25	200	3.3	8.4	8.6	5.1
1.25	1000	3.3	4.6	4.7	1.3
1.25	5000	3.3	3.5	3.5	0.3
1.50	200	5.8	9.2	9.6	3.4
1.50	1000	5.8	6.4	6.5	0.6
1.50	5000	5.8	5.9	5.9	0.1
2.00	200	9.3	11.2	11.7	1.9
2.00	1000	9.3	9.6	9.6	0.3
2.00	5000	9.3	9.3	9.3	0.0

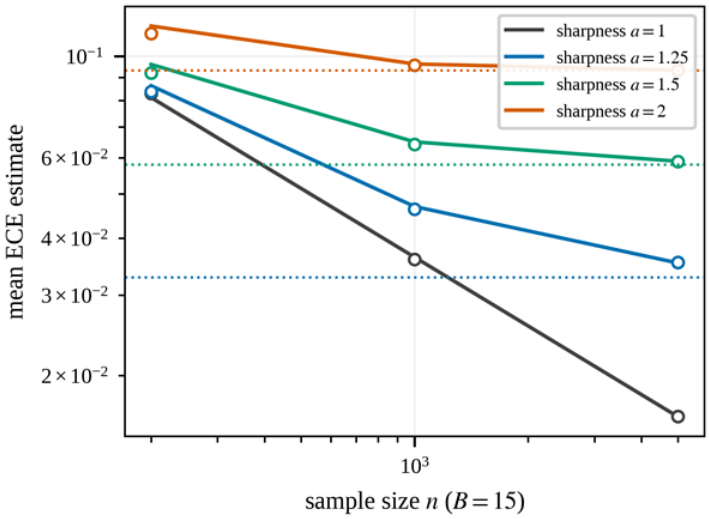


Figure 5: Mean ECE estimate for overconfident models of increasing sharpness a ($B = 15$). Lines: Eq. (2). Circles: simulation. Dotted: the true calibration error of each model.

6.3. How Many Cases Are Needed

Eq. (4) gives the sample size needed to bring the bias of a calibrated model below ε. With 10 bins and uniform scores it is 2,486 cases for ε = 0.02 and 9,945 for ε = 0.01; with 20 bins 4,932 and 19,730. The simulation agrees within 6% (Table 3 and Figure 7). The requirement is large: to say that a perfectly calibrated model has an ECE below one point needs about ten thousand cases, and the smaller the true error, the more cases are needed to see that it is small.

Table 3: Sample size for a mean ECE of at most ε for a calibrated model with uniform scores: Eq. (4) and simulation.

B	ε	Eq. (4)	Simulated
10	0.02	2,486	2,528
10	0.01	9,945	10,316
20	0.02	4,932	4,726
20	0.01	19,730	20,868

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

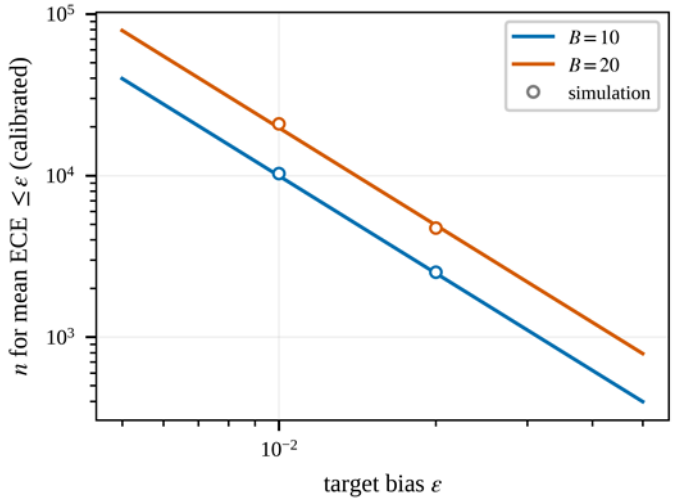


Figure 7: Sample size needed for the mean ECE of a calibrated model with uniform scores to fall below ε. Lines: Eq. (4). Circles: simulation.

6.4. The number of bins

The number of bins trades two errors. Figure 6a shows the bias of the estimate against the true calibration error for a = 1.5. With few bins the estimate is below the truth, since errors of opposite sign cancel inside a bin: at 1,000 cases and 3 bins the bias is -0.7 points. With many bins the noise dominates: 4.4 points at 100 bins. Eq. (2) tracks both. The root mean squared error (Figure 6b) is smallest at 3 bins for 300 cases (RMSE 1.8 points), 5 bins for 1,000 cases (1.2), and 12 bins for 5,000 cases (0.5). The best number grows with n, but slowly, and the error is flat over a range of bins; a common choice of 10 to 15 bins is reasonable above about 1,000 cases and too many below.

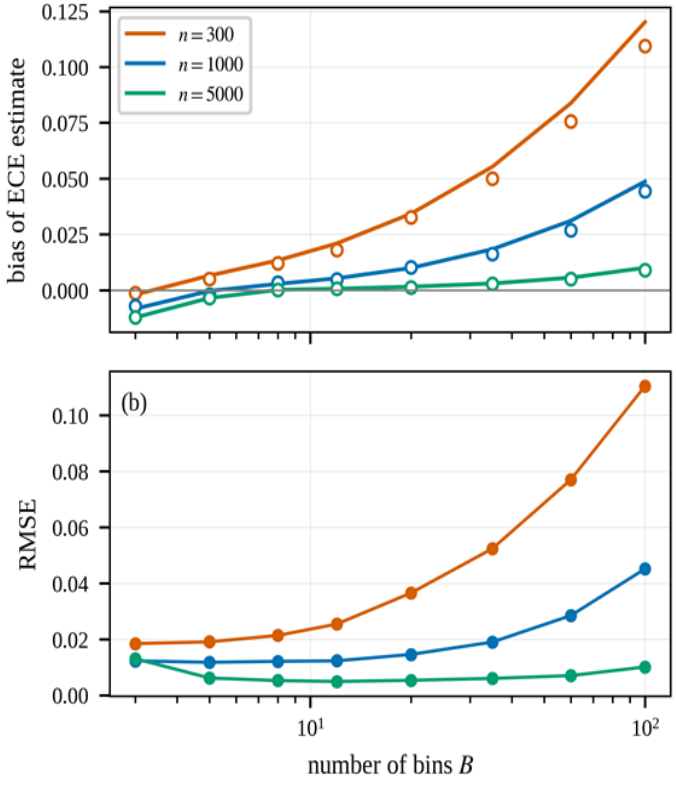


Figure. 6 The number of bins at a = 1.5. (a) Bias of the ECE estimate against the true calibration error: Eq. (2) (lines) and simulation (circles). (b) Root mean squared error of the estimate (300 repetitions).

6.5. Equal-width bins

Equal-width bins are the other common choice. Eq. (5) matches the simulation within 7.8% (Table 7 and Figure 10). The two binning schemes give similar biases for uniform scores (3.1% and 3.1% at 1,000 cases, 10 bins) but differ when the scores are concentrated. For low-risk scores, equal-mass bins give 2.5% and equal-width bins 2.0%; for bimodal scores the order is reversed, 2.0% for equal-mass and 2.4% for equal-width. A reported ECE therefore depends on the binning scheme as well as on the sample size, and comparisons should fix both.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

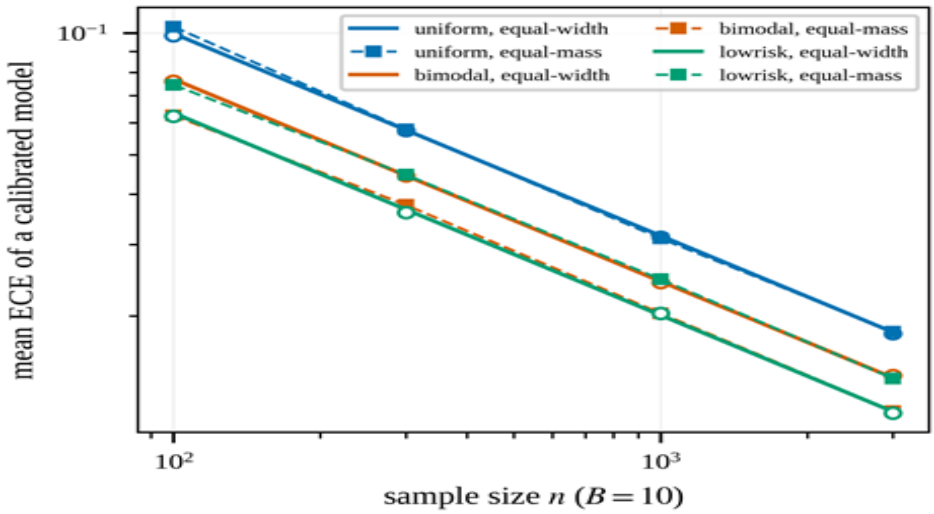


Figure. 10 Mean ECE of a calibrated model with equal-width (lines: Eq. 5, circles: simulation) and equal-mass (squares) bins, $B = 10$.

6.6. Can The Noise Be Removed?

Table 4 and Figure 8 compare four estimators. At $n = 1{,}000$ and $B = 15$, the plain estimate of a calibrated model has bias 3.6 points. The null correction removes it (bias 0.3) and the soft threshold nearly does (0.8), but for a model with $a = 1.5$ and true error 5.8% they understate it by 2.9 and 3.1 points, a worse error than the plain estimate (0.7). The reason is that the correction subtracts the noise of a calibrated model from bins that are not calibrated, where the noise has already been partly absorbed into the gap.

The inversion estimator solves for the true gap in each bin: it is nearly unbiased for $a = 1.5$ (-0.1), with RMSE 1.2 points, about the same as the plain estimate (1.2); it is clearly better at $a = 1.25$ (1.2 against 1.7) and slightly worse at $a = 2$ (1.3 against 1.2); for a calibrated model its bias is 2.2, which is much worse than the null-corrected one. No estimator is best everywhere. At 200 cases the same pattern holds with larger errors.

Table. 4 Mean ECE (%) of a calibrated model with equal-width bins ($B = 10$): simulation (500 repetitions), Eq. (5), and equal-mass simulation.

Scores	n	Equal-width sim.	Eq. (5)	Equal-mass sim
uniform	100	9.9	10.0	10.4
uniform	1000	3.1	3.2	3.1
uniform	3000	1.8	1.8	1.8
bimodal	100	7.6	7.7	6.3
bimodal	1000	2.4	2.4	2.0
bimodal	3000	1.4	1.4	1.2
low-risk	100	6.2	6.3	7.4
low-risk	1000	2.0	2.0	2.5
low-risk	3000	1.2	1.2	1.4

Table. 5 Bias (RMSE) in percentage points of four ECE estimators, a = model sharpness, true error in the second column.

n, B	a	True (%)	Plain	Null Corr	Soft thr	Inversion
1000, 15	1.00	0.0	3.6 (3.6)	0.3 (0.5)	0.8 (0.9)	2.2 (2.4)
1000, 15	1.25	3.3	1.4 (1.7)	-2.2 (2.3)	-1.9 (2.0)	0.2 (1.2)
1000, 15	1.50	5.8	0.7 (1.2)	-2.9 (3.1)	-3.1 (3.2)	-0.1 (1.2)
1000, 15	2.00	9.3	0.3 (1.2)	-3.3 (3.5)	-3.9 (4.0)	-0.1 (1.3)
200, 10	1.00	0.0	6.6 (6.8)	0.8 (1.4)	1.8 (2.1)	4.1 (4.5)
200, 10	1.25	3.3	3.8 (4.2)	-2.2 (2.6)	-1.5 (1.9)	1.4 (2.5)
200, 10	1.50	5.8	2.5 (3.2)	-3.7 (4.1)	-3.4 (3.7)	0.3 (2.4)
200, 10	2.00	9.3	1.1 (2.7)	-5.2 (5.6)	-5.6 (5.8)	-0.5 (2.8)

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

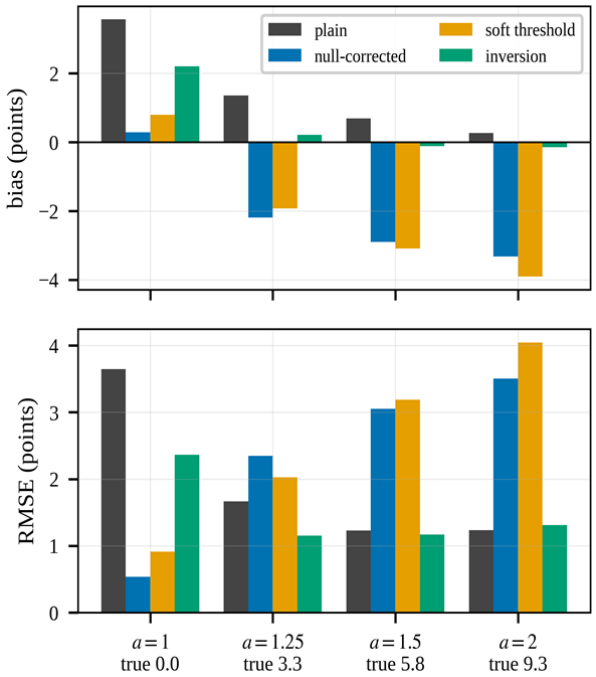


Figure. 8 Four estimators of the ECE at n = 1,000, B = 15 for models of increasing sharpness a (800 repetitions). (a) Bias. (b) RMSE. The true value (points) is shown under each group.

6.7. Real Data

Table 5 and Figure 9 show five models. The logistic regression models on all features are nearly calibrated and confident: their estimates (1.0% and 0.9% at 400 cases) are close to the calibrated floor (1.2% and 0.9%), and the floor is small because the predictions are near 0 and 1. The model with two weak features has moderate probabilities and a large floor. Its estimate is 16.2% at 50 cases and 6.5% at 400, against 6.0% on all 569 cases, and almost all of the estimate at small n is noise: the floor is 15.7% at 50 cases. The naive Bayes models are overconfident. Their estimates (3.7% on breast cancer and 29.3% on digits) are far above the floors (0.3% and 0.4%) and are stable across sample sizes, and for digits the estimate (29.1% at 50 cases) is already at its final value. The floor is a useful diagnostic: when the estimate is close to it, the data cannot distinguish the model from a calibrated one.

Table.6 Real data (B = 10, 1,000 subsamples): mean ECE estimate (%) and the calibrated floor at n = 50 and 400, and the ECE of all cases.

Model	All cases	n = 50 ECE	n = 50 floor	n = 400 ECE	n = 400 floor
breast, logistic	1.0	2.8	3.5	1.0	1.2
breast, 2 weak features	6.0	16.2	15.7	6.5	5.6
breast, naive Bayes	3.5	5.1	0.9	3.7	0.3
digits 9, logistic	0.9	1.7	2.4	0.9	0.9
digits 9, naive Bayes	29.2	29.1	1.0	29.3	0.4

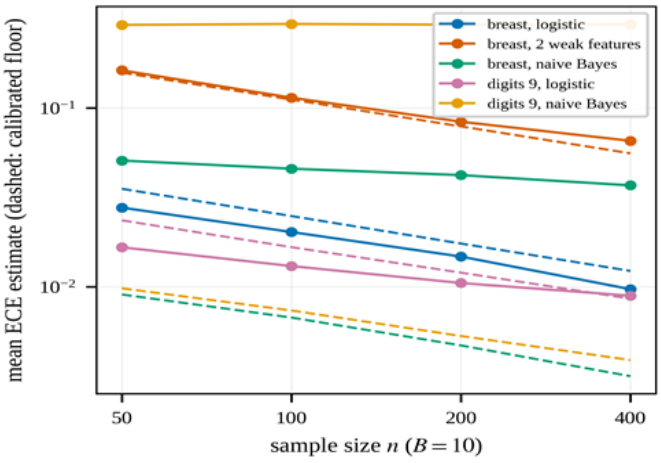


Figure. 9 Real data: mean ECE estimate against sample size (solid) and the value expected if the model were perfectly calibrated (dashed), computed from the scores of each subsample.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

6.8. A planning table

Table 6 evaluates Eq. (4) for the number of cases needed to keep the bias of a calibrated model below ε, for uniform and low-risk scores. The low-risk case, where most cases have small probabilities, needs fewer cases because $\bar{s}$ is smaller, 0.30 against 0.40. With 10 bins and ε = 0.02, the requirement is 2,486 cases for uniform scores and 1,474 for low-risk scores.

Table 6: Cases needed (Eq. 4) for a mean ECE of at most ε for a calibrated model, by number of bins.

B	Uniform ε = 0.05	Uniform ε = 0.02	Uniform ε = 0.01	Low-risk ε = 0.02	Low-risk ε = 0.01
5	203	1,266	5,063	744	2,976
10	398	2,486	9,945	1,474	5,896
15	593	3,709	14,835	2,206	8,824
20	789	4,932	19,730	2,938	11,753
50	1,965	12,283	49,134	7,336	29,344

6.9. A Practical Recipe

The results suggest a short procedure. First, never report an ECE without its sample size and number of bins: the same model can have an estimate of several points or less than one. Second, compute the calibrated floor from the predicted probabilities, $\sqrt{(2/\pi)}$ × the mean of $\sqrt{(\text{mean } q(1-q)/n_b)}$, and report it with the estimate; if the estimate is close to the floor, the evidence for miscalibration is weak. Third, use Eq. (4) or Table 6 to plan: expect thousands of cases to make the bias of a calibrated model small. Fourth, choose few bins when the sample is small (5 to 10 below about 1,000 cases) and more when it is large. Fifth, do not subtract the noise without checking: a correction that works for a calibrated model can understate the error of a miscalibrated one; if a correction is needed, the inversion estimator reduces the bias when the model is known to be miscalibrated, at a similar RMSE.

Discussion and Limitations

What we think carries over. Three points seem general. A binned absolute-value estimator is biased upwards by an amount that depends on the sample size and the number of bins, and that can be as large as the quantity being measured. The bias is largest for the models that are closest to calibrated, which are the ones people most want to certify. And any comparison of calibration between models or methods should use the same sample size and number of bins, since otherwise the comparison measures the noise.

Synthetic setting. Eq. (2) uses a normal approximation for each bin. It was accurate in our settings with at least 20 cases per bin, and we did not study smaller bins, where the approximation breaks down. We studied binary classification and the ℓ1 error with equal-mass and equal-width bins; multi-class confidence calibration and the ℓ2 error have different formulas. The miscalibration family has one parameter and is monotone.

Real data. We used two small datasets and cross-validated probabilities; the "truth" for the real data is the ECE on all cases, which has its own noise, and we did not have a ground-truth calibration error. The floor is a diagnostic and not a test. We did not study the other corrections in the literature, such as debiased ℓ2 estimators [5], kernel and smoothing estimators, or bootstrap bias correction.

7. Conclusion

The binned ECE of a perfectly calibrated model is positive, with expected value $\bar{s}\sqrt{(2B/(\pi n))}$, which is 3.8% at 1,000 cases and 15 bins. The general case is a sum of folded-normal means that matched simulation within 5.3%. Keeping the bias of a calibrated model below one point needs about 9,945 cases with 10 bins, a number that grows with the number of bins and with the inverse square of the target. The number of bins that minimises the error is small and grows slowly with n, and noise corrections must be used with care. Next steps are to extend the analysis to multi-class calibration, equal-width bins and the ℓ2 error, and to confidence intervals for the ECE that take the bias into account.

References

[1]. M. P. Naeini, G. F. Cooper, M. Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning into quantiles. In *AAAI*, 2015.

[2]. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[3]. A. Niculescu-Mizil, R. Caruana. Predicting good probabilities with supervised learning. In *ICML*, 2005.

[4]. M. H. DeGroot, S. E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32(1/2):12–22, 1983.

[5]. A. Kumar, P. Liang, T. Ma. Verified uncertainty calibration. In *NeurIPS*, 2019.

[6]. J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, D. Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019.

[7]. J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, T. B. Schön. Evaluating model calibration in

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P6

classification. In *AISTATS*, 2019.

[8]. T. Gneiting, A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.*, 102(477):359–378, 2007.

[9]. G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[10]. J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, MIT Press, 1999.

[11]. B. Zadrozny, C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *ICML*, 2001.

[12]. F. C. Leone, L. S. Nelson, R. B. Nottingham. The folded normal distribution. *Technometrics*, 3(4):543–550, 1961.

[13]. F. Pedregosa et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[14]. W. N. Street, W. H. Wolberg, O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *SPIE*, vol. 1905, 1993.

[15]. E. Alpaydin, C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

[16]. D. Widmann, F. Lindsten, D. Zachariah. Calibration tests in multi-class classification: a unifying framework. In *NeurIPS*, 2019.

[17]. J. Bröcker. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate Dynamics*, 39:655–667, 2012.

A Derivations

Eq. (2). Given the predicted probabilities, the bin difference $\bar{y}_b - \bar{q}_b$ is a sum of independent terms $(y_i - q_i)/n_b$ with mean $(p_i - q_i)/n_b$ and variance $p_i(1 - p_i)/n_b^2$. Averaging gives mean δ_b and variance σ_b^2, and by the central limit theorem it is approximately normal. The mean of the absolute value of $N(\delta, \sigma^2)$ is $\sigma\sqrt{(2/\pi)}\, e^{-\delta^2/(2\sigma^2)} + \delta(1 - 2\Phi(-\delta/\sigma))$ [12]. Weighting by w_b gives Eq. (2).

Eq. (3). At $\delta = 0$, $F = \sigma\sqrt{(2/\pi)}$. With $n_b = n/B$, $\sigma_b = \sqrt{(v_b B/n)}$, and averaging over B bins with weights $1/B$ gives $\bar{s}\sqrt{(2B/(\pi n))}$.

Eq. (4). Setting Eq. (3) equal to ε and solving for n gives $n = (2B/\pi)(\bar{s}/\varepsilon)^2$.

Inequalities. $F(\delta, \sigma) \geq |\delta|$ by Jensen's inequality, and $F(\delta, \sigma) - |\delta| = E|\delta + \sigma Z| - |\delta| \leq E|\sigma Z| = \sigma\sqrt{(2/\pi)}$ by the triangle inequality.

B Parameters

Calibrated: p uniform, Beta(0.3, 0.3) or Beta(1, 6), $q = p$, $n \in \{100, 300, 1{,}000, 3{,}000, 10{,}000\}$, $B \in \{5, 10, 15, 20, 50\}$, 600 repetitions, equal-mass bins. Miscalibrated: $z \sim N(0, 2^2)$, $q = \text{logistic}(a\, z)$, $a \in \{1, 1.25, 1.5, 2\}$; true error from 10^6 cases. Required-n search on 14 sample sizes from 100 to 60,000. Bin study: $a = 1.5$, $B \in \{3, 5, 8, 12, 20, 35, 60, 100\}$, $n \in \{300, 1{,}000, 5{,}000\}$, 300 repetitions. Estimator study: 800 repetitions. Inversion solved by bisection on F. Real data: five-fold cross-validated probabilities from logistic regression ($C = 1$, standardised features) and Gaussian naive Bayes; subsamples without replacement of 50, 100, 200 and 400 cases, 1,000 repetitions, $B = 10$. All figures were produced by code at 600 dpi and are printed at their stored width. Code, results and figures are in the supplementary archive.

How To Cite:

Mesfin Dagne , " Analytical Characterization of Finite-Sample Bias in Expected Calibration Error for Probabilistic Classifiers using Artificial Intelligence ", International Journal of Research and Development in Engineering Sciences (IJRDES), Volume 8 , Issue 6 , pp 58 – 67 , November 2026.
CrossRef DOI: https://doi.org/10.63328/ijrdes-v8ri6p6

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com