Calibrating the Calibration Tester: Optimal Binning and Minimax Calibration Testing for Continuous Predictive Models

Alon Kipnis
School of Computer Science, Reichman University, Herzliya, Israel

Abstract

Evaluating the calibration of continuous predictive models relies heavily on the binned Probability Integral Transform (PIT). Practitioners routinely use arbitrary bin counts (N) and standard χ^2 goodness-of-fit tests, which lack theoretical guarantees against worst-case calibration errors, especially in the large-N regime or when the sample size fluctuates. In this work, we translate recent advances in minimax uniformity testing to the machine learning calibration setting. By mapping regression Expected Calibration Error (ECE) to the ℓ_p distance from uniformity, we provide two rigorous tools for practitioners: (1) a formula for the maximum allowable bins $N_{\max}$ to guarantee detection of a target ECE, and (2) a test that achieves minimax optimal rates. Finally, we discuss the trade-off between discretization bias and statistical noise, and demonstrate how the formula for $N_{\max}$ provides a principled way to choose N that balances these two effects.

1 Introduction

Continuous calibration evaluates whether a model's predictive distribution assigns probabilities that agree with the empirical frequencies of observed outcomes. In the continuous setting, this is typically assessed using the Probability Integral Transform (PIT) together with histogram binning, where the number of bins N is usually treated as a heuristic design choice rather than a principled parameter. Yet the choice of this resolution parameter masks a fundamental statistical trade-off. If N is too small, complex or oscillatory miscalibrations may cancel within wide bins, falsely suggesting calibration. If N is too large, the finite test set of n samples is spread too thin, weakening the statistical signal and reducing test power; see discussions in Wallis (2003); Nixon et al. (2019); Lee et al. (2023) and references therein. Existing approaches for choosing N rely primarily on data-driven procedures that require additional calibration of the null distribution. For example, Lee et al. (2023) proposed to select N by estimating a smoothness parameter from the data.

In this work, we eliminate this arbitrary hyperparameter tuning by adapting recent sharp minimax bounds for uniformity testing into the machine learning calibration setting (Kipnis, 2025, 2026). We establish the fundamental statistical limits of binned calibration evaluation and provide practitioners with a rigorous formula for the maximum allowable resolution $N_{\max}$ that still guarantees detection of a target ECE against ℓ_p alternatives with a separation parameter ϵ. We further introduce a variance-stabilized test statistic that achieves asymptotically minimax detection rates against ℓ_p-ECE departures, while remaining easy to calibrate and robust to random test set sizes, unlike the standard χ^2 statistic. As is well-known, binning the PIT histogram smooths the underlying density, so fluctuating and oscillatory miscalibrations cancel within each bin and become invisible to any binned test (cf. Roelofs et al. (2022)). This reveals $N_{\max}$ as the resolution that optimally trades discretization bias against statistical noise. We illustrate the full picture with a Monte Carlo experiment on a highly oscillating density.

1.1 Outline

In Section 2, we formulate the calibration testing problem and discuss the smoothing effect of binning. In Section 3, we introduce the robust asymptotically minimax testing procedure and its risk. In Section 4 we derive a formula for the maximum number of bins $N_{\max}$ given ϵ, n, and target risk R. In Section 5, we illustrate the theory with a numerical example. Finally, in

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco. Copyright 2026 by the author(s).

Section 6 we conclude.

2 Problem Formulation

2.1 Continuous Calibration and PIT

Let $(X, Y) \sim P_{X,Y}$ be a random vector where $X \in \mathcal{X}$ and $Y \in \mathbb{R}$. Given a trained predictive model that outputs a conditional Cumulative Distribution Function (CDF) $\hat{F}(y|x)$, we are interested in testing the hypothesis that the model is perfectly calibrated on a test set $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, assumed to be i.i.d. from $P_{X,Y}$.

In practice, this is done via the PIT (Rosenblatt, 1952; Dawid, 1984; Diebold et al., 1998). The PIT value for an observation (X_i, Y_i) is defined as: $V_i = \hat{F}(Y_i|X_i)$. Under the null hypothesis of perfect distributional calibration, $H_0 : V_i \sim \text{Unif}(0, 1)$ for all $i = 1, \ldots, n$.

Let $Q(v) := \Pr[V_1 \leq v]$ be the true CDF of V_i and let $q(v)$ be the true density, which we assume to exist. We quantify miscalibration using the Expected Calibration Error (ECE) (Naeini et al., 2015; Chung et al., 2021):

$$\text{ECE}_p(Q) := \left(\int_0^1 |q(v) - 1|^p dv \right)^{1/p}. \tag{1}$$

For $p = 1$, this is the area between the two curves in a standard reliability diagram (cf. Naeini et al. (2015); Guo et al. (2017)).

2.2 Binning and Discretized ECE

In practice, validating H_0 involves binning the PIT values into N equal-width bins with edges $0 = t_0 < t_1 < \ldots < t_N = 1$ and comparing the binned counts

$$Z_j := \#\{i : V_i \in [t_{j-1}, t_j]\}, \quad j = 1, \ldots, N,$$

to the uniform baseline (cf. Kuleshov et al. (2018); Song et al. (2019); Levi et al. (2022)).

Let $\mathbf{q} := (q_1, \ldots, q_N)$ be the true probability mass function of $\{Z_j, j = 1, \ldots, N\}$. Namely, $q_j := Q(t_j) - Q(t_{j-1})$. Let $\mathbf{u} := (1/N, \ldots, 1/N)$ be the uniform baseline. Define the probability model

$$H(\mathbf{q}) : (Z_1, \ldots, Z_N) \sim \text{Multi}(n, \mathbf{q}). \tag{2}$$

Subject to this discretization, the null hypothesis becomes $H_0^{(N)} = H(\mathbf{u})$ and the ECE is approximated by the discretized ECE (a.k.a. the binned ECE in Roelofs et al. (2022)):

$$\text{ECE}_p^{(N)}(\mathbf{q}) := \left(\sum_{j=1}^N |q_j - 1/N|^p \right)^{1/p}.$$

2.3 Smoothing Effect of Binning

Notice that, for $p \geq 1$, we have

$$\left(\text{ECE}_p^{(N)}(\mathbf{q})\right)^p = \sum_{j=1}^N \left| \int_{t_{j-1}}^{t_j} (q(v) - 1) dv \right|^p \tag{3}$$

$$\leq \sum_{j=1}^N \int_{t_{j-1}}^{t_j} |q(v) - 1|^p \, dv = \text{ECE}_p(Q)^p, \tag{4}$$

so $\text{ECE}_p^{(N)}(\mathbf{q}) \leq \text{ECE}_p(Q)$. The transition between the lines is due to Jensen's inequality and convexity of $x \to x^p$. The key lesson from this inequality is that bin averaging smooths fluctuations in $q(v)$, making it look closer to uniform than the true distribution Q. This smoothing effect is the penalty in taking N to be too small, whereas the focus of this paper is on the penalty in taking N to be too large. We discuss this trade-off in more detail in Sections 4 and 5 below.

2.4 Minimax Formulation

A test ψ for binned calibration is a function from $\{Z_j : j = 1, \ldots, N\}$ to $\{0, 1\}$ whose value represents whether $H_0^{(N)}$ is rejected. The risk of ψ against a simple alternative $H(\mathbf{q})$ is defined as:

$$R(\psi; \mathbf{q}) := \Pr\left[\psi(Z_1, \ldots, Z_N) = 1 \mid H_0^{(N)}\right] + \Pr\left[\psi(Z_1, \ldots, Z_N) = 0 \mid H(\mathbf{q})\right].$$

The minimax risk against discretized ECE alternatives with separation parameter ϵ is defined as:

$$R_N^* := \inf_{\psi} \sup_{\mathbf{q} \,:\, \text{ECE}_p^{(N)}(\mathbf{q}) \geq \epsilon} R(\psi; \mathbf{q}).$$

Our goal in this paper is to characterize R_N^* and the test attaining it, and understand how R_N^* is affected by the number of bins N. We achieve this characterization in an asymptotic regime, where n and N tend to infinity while ϵ typically tends to zero.

3 Robust Minimax Testing under ECE Departures

3.1 Asymptotically Minimax Test

Let

$$\xi_{n,N,\epsilon} := \frac{\epsilon^2 n}{\sqrt{2} N^{2/p - 3/2}}. \tag{5}$$

For a reason that will become clear soon, we call $\xi_{n,N,\epsilon}$ the signal-to-noise ratio (SNR) parameter of the testing

problem. In the important case of $p = 1$ we have $\xi_{n,N,\epsilon} = n\epsilon^2/\sqrt{2N}$.

Consider the test statistic

$$T_N := \sqrt{\frac{N}{2n^2}} \sum_{j=1}^{N} \left[(Z_j - n/N)^2 - Z_j \right], \tag{6}$$

and a test ψ^* that rejects H_0 if

$$T_N > z_{1-\alpha^*}, \quad \alpha^* = \Phi(-\xi_{n,N,\epsilon}/2), \tag{7}$$

where Φ is the standard normal CDF and z_t is the t-th quantile of the standard normal distribution. The following result is a direct consequence of Kipnis (2025) and Kipnis (2026), when adjusted for the notation introduced in Section 2 above.

Theorem 3.1. *Consider an asymptotic setting where $N = o(n^2)$ and $\xi_{n,N,\epsilon} \to \xi^*$ for some $\xi^* \in (0, \infty)$, $p \in (0, 2]$. Then*

(i) $R^*_N = 2\Phi(-\xi^*/2) + o(1)$.

(ii) For the test ψ^* defined in (7)*, we have*

$$\sup_{\mathbf{q}\,:\,\mathrm{ECE}_p^{(N)}(\mathbf{q}) \geq \epsilon} R(\psi^*; \mathbf{q}) = R^*_N + o(1).$$

Namely, in the asymptotic setting described above, the asymptotic minimax risk is $2\Phi(-\xi^*/2)$ and the test defined in (7) is asymptotically minimax.

Furthermore, suppose that the sample size n in (2) is random with $n \sim \mathrm{Pois}(M)$, but the rest of the setting is unchanged[1]. Then (i) and (ii) in Theorem 3.1 hold with M replacing n. In particular, the test (7) achieves the minimax risk R^*_N against ECE departures with separation parameter ϵ under the random sample size model.

3.2 Testing based on the χ^2 statistic

The χ^2 statistic is defined as:

$$T_{\chi^2} := \sum_{j=1}^{N} \frac{(Z_j - n/N)^2}{n/N}. \tag{8}$$

It follows from Kipnis (2025) that when the sample size n is random with $n \sim \mathrm{Pois}(M)$ and under the same conditions as in Theorem 3.1, the maximal risk under the alternative of the best-calibrated test based on χ^2 is $2\Phi(-\xi_{\chi^2}/2) + o(1)$, where $\xi_{\chi^2} = \xi_{n,N,\epsilon}\sqrt{M/(M+N)}$. Therefore, this test is strictly sub-optimal compared to (7) unless $N/M \to 0$.

[1]This randomness changes the multinomial distribution in (2) to a multivariate Poisson distribution, which aligns with the setting of Kipnis (2025).

4 Determining the Maximum Number of Bins

The expression for R^*_N from Theorem 3.1 yields the maximum number of bins $N_{\max}$ for a given separation parameter ϵ, sample size n, and a target testing risk R.

$$N_{\max} = \left\lfloor \left(\frac{n^2\epsilon^4}{8[\Phi^{-1}(1 - R/2)]^2} \right)^{p/(4-3p)} \right\rfloor. \tag{9}$$

We note that if one wishes to control the Type I error at α instead of the risk R, the worst-case Type II error of the test (7) (with $z_{1-\alpha}$ replacing $z_{1-\alpha^*}$) is

$$\beta^*(\alpha) = \Phi(z_{1-\alpha} - \xi_{n,N,\epsilon}).$$

To guarantee $\beta^*(\alpha) \leq \beta$, one needs $\xi_{n,N,\epsilon} \geq z_{1-\alpha} + z_{1-\beta}$, giving

$$N_{\max} = \left\lfloor \left(\frac{n^2\epsilon^4}{2\left(z_{1-\alpha} + z_{1-\beta}\right)^2} \right)^{p/(4-3p)} \right\rfloor. \tag{10}$$

It is instructive to interpret formula (9) in light of Section 2.3. The choice of N is governed by two competing effects:

- **Discretization bias** (small N). By Section 2.3, binning smooths the density. If $q(v)$ fluctuates, the within-bin integrals partially or fully cancel, yielding $\mathrm{ECE}_p^{(N)}(\mathbf{q}) \ll \mathrm{ECE}_p(Q)$. The test then operates against a much weaker separation than the true continuous ECE.
- **Statistical noise** (large N). The SNR $\xi_{n,N,\epsilon}$ decays as $N^{3/2-2/p}$ (for $p = 1$, as $N^{-1/2}$), because each additional bin spreads the n samples thinner.

The formula $N_{\max}$ balances these two effects: it is the finest resolution at which the SNR $\xi_{n,N,\epsilon}$ still guarantees the target risk R. Using $N < N_{\max}$ may achieve the target risk against the *discretized* alternative, but the discretization bias may mask the true continuous miscalibration; see the example in the next section for a concrete illustration.

5 Numerical Example

We illustrate the discretization-versus-noise trade-off discussed in Section 4 with a Monte Carlo experiment.

5.1 Setup

We consider $n = 5000$ test samples from a regression model whose true conditional CDF is $F(y|x)$. A miscalibrated predictor applies an oscillatory perturbation

to its predicted CDF in the probability domain:

$$\hat{F}(y|x) = F(y|x) + \tfrac{a}{2\pi k}\sin\big(2\pi k \cdot F(y|x)\big), \qquad (11)$$

with $k = 50$ oscillation cycles and amplitude $a = 0.2$. This is a valid CDF for $|a| < 1$ and corresponds to a miscalibrated quantile map, the kind of error that can arise from an imperfect post-hoc recalibration step applied in the probability domain. The PIT values $V_i = \hat{F}(Y_i|X_i)$ have density

$$q(v) = 1 + a\cos(2\pi k v), \quad v \in [0,1], \qquad (12)$$

whose CDF $Q(v) = v + \frac{a}{2\pi k}\sin(2\pi k v)$ oscillates rapidly around the diagonal. The asymptotic (large-N) binned ECE is $\epsilon_\infty = 2a/\pi \approx 0.127$. We are interested in detecting this miscalibration with a test that achieves a target minimax risk against ℓ_1-ECE departures with separation parameter ϵ_∞.

Notice that due to the Shannon-Nyquist sampling theorem (Shannon, 1949), this setup has a natural critical sampling resolution $N_{\text{smp}} = 2k$. Below this resolution it is impossible to reconstruct $q(v)$ with vanishing error even as $n \to \infty$.

5.2 The bin-count trade-off

When the PIT values are discretized into N equal-width bins, a direct calculation of the absolute bin deviations as in (3) gives the approximate binned ECE

$$\text{ECE}_1^{(N)} \;\approx\; \epsilon_\infty \cdot |\text{sinc}(k/N)|\,, \qquad \text{sinc}(x) := \frac{\sin(\pi x)}{\pi x},$$

so the discretization attenuates the true ECE by a factor that grows smoothly from near zero (when $N \ll N_{\text{smp}}$) to one, reaching about 64% at $N = N_{\text{smp}} = 2k$. At the same time, the SNR $\xi_{n,N,\epsilon_\infty} = \epsilon_\infty^2 n/\sqrt{2N}$ decays as $N^{-1/2}$, so the testing risk eventually climbs back as N grows. The formula (9) with $p = 1$ yields $N_{\max} = 303$ at the target risk $R = 0.1$.

5.3 Monte Carlo protocol

For each value of N in a logarithmic grid from 5 to 5000, we repeat the following 1000 times: (i) draw n null PIT values $V_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$ and compute the standardized statistic T_N; (ii) draw n alternative PIT values from density q in (12) and compute the same statistic. For each N, we use the minimax test based on the true binned ECE $\epsilon^{(N)} := \text{ECE}_1^{(N)}(q)$. We report the resulting empirical risk (Type I + Type II).

5.4 Results

Figure 1 displays the central result. The risk curve exhibits the U-shape predicted by the theory: when

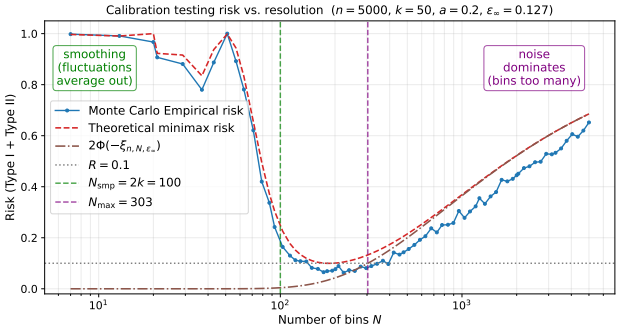

Figure 1: Testing risk (Type I + Type II) as a function of the number of bins N, averaged over 1000 Monte Carlo repetitions. Blue: empirical risk of the minimax test (7). Red dashed: theoretical minimax risk $2\Phi(-\xi_{n,N,\epsilon^{(N)}}/2)$, where $\epsilon^{(N)} = \text{ECE}_1^{(N)}(q)$ is the binned ECE. The critical sampling resolution $N_{\text{smp}} = 2k$ (green) and $N_{\max}$ of (9) (purple) are marked.

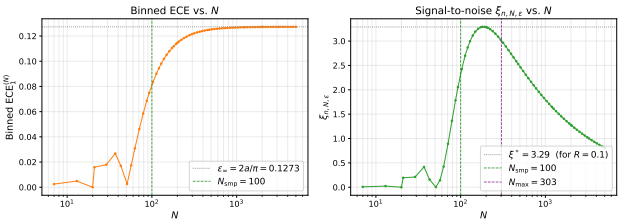

Figure 2: Left: binned $\text{ECE}_1^{(N)}$ as a function of N. Below $N_{\text{smp}} = 2k$, bin-averaging cancels the oscillation and the ECE is near zero. Right: the SNR parameter $\xi_{n,N,\epsilon^{(N)}}$, where $\epsilon^{(N)} = \text{ECE}_1^{(N)}(q)$. It peaks when the trade-off between increasing ECE and decreasing per-bin sample size is balanced, then decays as $N^{-1/2}$ for large N.

$N \ll N_{\text{smp}}$, bin-averaging drives $\text{ECE}_1^{(N)}$ to near zero and the test has no power. When $N \gg N_{\max}$, the binned ECE is fully resolved, but statistical noise dominates (n/N samples per bin) and the risk climbs back toward 1. The two competing effects are also visualized separately in Figure 2.

Figure 3 provides single-experiment snapshots at three representative resolutions, illustrating the smoothing regime (left), optimal detection regime (center), and noise-dominated regime (right).

6 Conclusion

We provided a rigorous statistical foundation for the empirical evaluation of continuous predictive models. By translating the machinery of minimax uniformity testing to the domain of regression calibration, we eliminated the heuristic guesswork traditionally associated

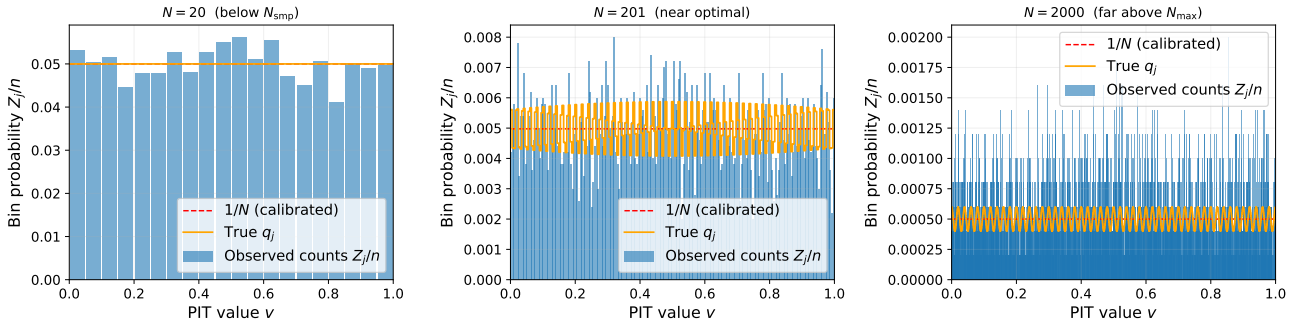

Figure 3: PIT histograms from a single experiment at three resolutions. Left ($N = 20$): bins are too wide; the oscillatory miscalibration (orange: true q_j) is indistinguishable from the uniform baseline (red dashed). Center ($N = 201$): the oscillation pattern starts to be resolved against the sampling noise (blue bars). Right ($N = 2000$): the expected count per bin is only $n/N = 2.5$, statistical sampling noise masks the signal.

with binning the Probability Integral Transform (PIT). We established a precise limit $N_{\max}$, ensuring practitioners can safely maximize their histogram resolution without drowning the Expected Calibration Error (ECE) signal in multinomial noise. A key insight is that binning smooths the density, so $N_{\max}$ is the finest resolution at which this discretization bias is controlled while the statistical noise remains manageable.

Acknowledgments

This work was supported by the US-Israel Binational Science Foundation (BSF) under grant 2022124.

References

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. (2021). Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:10971–10984.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Kipnis, A. (2025). The minimax risk in testing uniformity of poisson data under missing ball alternatives. *IEEE Transactions on Information Theory*.

Kipnis, A. (2026). Sharp constant minimax risk in uniformity testing via a central limit theorem in a compound Poisson setting. in preparation.

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning (ICML)*, pages 2796–2804. PMLR.

Lee, D., Huang, X., Hassani, H., and Dobriban, E. (2023). T-cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335):1–72.

Levi, D., Gispan, L., Giladi, N., and Fetaya, E. (2022). Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Blei, D. (2019). Measuring calibration in deep learning. In *CVPR Workshops*, volume 2.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. (2022). Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4036–4054. PMLR.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

Shannon, C. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Song, H., Diethe, T., Kull, M., and Flach, P. (2019). Distribution calibration for regression. In *International Conference on Machine Learning (ICML)*, pages 5897–5906. PMLR.

Wallis, K. F. (2003). Chi-squared tests of interval and density forecasts, and the bank of england's fan charts. *International Journal of Forecasting*, 19(2):165–175.