Scoring Methodology for Confidence Interval Estimation Challenge

Johan Persson
johan162@gmail.com

November 5, 2025

Abstract

We present a comprehensive scoring system for evaluating the quality of subjective confidence interval (CI) estimates in the context of commute time prediction. The methodology employs a multi-component penalty framework that simultaneously optimizes for interval precision, coverage accuracy, confidence calibration, and distributional balance. Drawing inspiration from Brier scoring principles, our approach extends traditional binary forecast evaluation to continuous interval estimation, incorporating quadratic penalties for deviations from target specifications. The system is designed to discourage gaming strategies while rewarding honest self-assessment and well-calibrated predictions. We demonstrate that the proposed scoring function effectively balances competing objectives through weighted penalty terms, with specific emphasis on ensuring both aggregate coverage (90% target) and tail symmetry (5% in each tail).

1 Introduction

1.1 Problem Statement

Accurate estimation of confidence intervals represents a fundamental challenge in probabilistic forecasting. In practical applications such as commute time prediction, users must balance competing objectives: intervals that are too narrow fail to capture sufficient observations, while excessively wide intervals, though technically correct, provide limited actionable information. Traditional scoring rules for probabilistic forecasts, such as the Brier score for binary events, do not directly extend to interval estimation problems with multiple quality dimensions.

1.2 Challenge Specification

Participants are asked to estimate a 90% confidence interval for their commute duration based on historical observations. Specifically, given a dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ of $n \geq 20$ observed commute times, participants must specify:

- **Interval bounds:** $[L, H]$ where L represents the lower bound and H the upper bound (in seconds)
- **Confidence level:** $c \in [5, 10]$ representing the participant's subjective confidence that their interval represents a true 90% CI

The target specification is a 90% confidence interval, corresponding to $\alpha = 0.10$ with $\alpha/2 = 0.05$ in each tail under a symmetric distribution assumption (z-score ± 1.645).

2 Theoretical Foundation

2.1 Brier Score and Mean Squared Error

The Brier score, introduced by Glenn W. Brier (1950), evaluates probabilistic forecasts for binary events through mean squared error:

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2$$

where f_i is the forecast probability and $o_i \in \{0, 1\}$ is the observed outcome. This quadratic penalty structure has desirable theoretical properties: it is strictly proper (encourages honest reporting), decomposable into calibration and refinement components, and provides smooth gradients that appropriately weight larger errors more heavily than smaller ones.

2.2 Extension to Interval Estimation

Our scoring methodology adapts the quadratic penalty principle to interval estimation by decomposing the problem into five measurable quality dimensions, each contributing a penalty term to the final score. The total score S is expressed as:

$$S = P_{\text{precision}} + P_{\text{miss}} + P_{\text{over}} + P_{\text{calib}} + P_{\text{balance}}$$

where lower scores indicate superior performance. This additive structure allows independent tuning of penalty weights while maintaining interpretability.

3 Scoring Components

3.1 Precision Penalty

Definition:

Let the sorted dataset be $\mathcal{D}_{\text{sorted}} = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Define the sample percentiles: - $p_5 = x_{(\lfloor 0.05n \rfloor)}$ (5th percentile) - $p_{95} = x_{(\lfloor 0.95n \rfloor)}$ (95th percentile)

Then:

$$P_{\text{precision}} = (L - p_5)^2 + (H - p_{95})^2$$

Rationale: The precision penalty measures how accurately the participant estimates the empirical 90% confidence interval boundaries. By penalizing squared deviations from the actual sample percentiles, this component rewards accurate boundary estimation regardless of the inherent variability of the commute times. This ensures fairness: participants with highly variable commutes (wide true intervals) are not penalized relative to those with consistent commutes (narrow true intervals). The penalty measures *orecasting skill*, not the characteristics of the underlying data.

Properties: - Quadratic in estimation error (Brier-inspired) - Scale-independent: measures relative accuracy - Fair across different commute variability levels - Zero penalty for perfect percentile matching

3.2 Miss Penalty

Definition: For each observation $x_i \in \mathcal{D}$, define the miss function:

$$m(x_i) = \begin{cases} (L - x_i)^2 & \text{if } x_i < L \\ (x_i - H)^2 & \text{if } x_i > H \\ 0 & \text{if } L \leq x_i \leq H \end{cases}$$

Let $\mathcal{M} = \{m(x_i) : m(x_i) > 0\}$ be the sorted set of non-zero misses, and $\mathcal{M}' = \mathcal{M}_{[\lfloor 0.1|\mathcal{M}| \rfloor :]}$ be $\mathcal{M}$ with the smallest 10% removed. Then:

$$P_{\text{miss}} = \frac{1}{n}\sum_{m \in \mathcal{M}'} m$$

Rationale: The squared distance formulation borrows directly from Brier's mean squared error principle, penalizing outliers more severely than near-misses. The 10% trimming provides tolerance for the 90% CI specification, effectively ignoring the smallest violations that fall within expected coverage gaps.

Properties: - Quadratic in miss distance (Brier-inspired) - Normalized by total sample size n - Robust to extreme outliers through 10% trimming - Invariant to confidence level (separated from calibration)

3.3 Overcoverage Penalty

Definition: Let $C_{\text{actual}} = \frac{100}{n}|\{x_i : L \leq x_i \leq H\}|$ be the empirical coverage percentage. With target coverage $C_{\text{target}} = 90$:

$$P_{\text{over}} = \begin{cases} 0.5 \cdot (H - L) \cdot \left(\frac{C_{\text{actual}} - 90}{10}\right)^2 & \text{if } C_{\text{actual}} > 95 \\ 0 & \text{otherwise} \end{cases}$$

Rationale: This component prevents a trivial gaming strategy where participants select arbitrarily wide intervals to guarantee coverage. The penalty activates only for coverage exceeding 95% (allowing 5% tolerance), and scales quadratically with both excess coverage and interval width. The multiplicative width term ensures that wasteful intervals incur proportionally higher costs.

Properties: - Activates beyond 5% tolerance threshold - Quadratic in excess coverage - Proportional to interval width - Coefficient 0.5 balances impact relative to other penalties

3.4 Confidence Calibration Penalty

Definition: Define:

$$P_{\text{below}} = \frac{100}{n}|\{x_i : x_i < L\}| \tag{1}$$

$$P_{\text{above}} = \frac{100}{n}|\{x_i : x_i > H\}| \tag{2}$$

$$P_{\text{outside}} = P_{\text{below}} + P_{\text{above}} \tag{3}$$

$$\Delta_{\text{cov}} = |P_{\text{outside}} - 10| \tag{4}$$

Map coverage deviation to ideal confidence:

$$c_{\text{ideal}} = \max\left(5, 10 - \frac{\Delta_{\text{cov}}}{10}\right)$$

Let $\Delta_c = |c - c_{\text{ideal}}|$ be the confidence mismatch. For $c = 10$ (claiming perfection), define tail imbalance:

$$\Delta_{\text{tail}} = |P_{\text{below}} - 5| + |P_{\text{above}} - 5|$$

Then:

$$P_{\text{calib}} = \begin{cases} 3.0 \cdot (H - L) \cdot \left(\frac{\Delta_{\text{cov}} + \Delta_{\text{tail}}}{5}\right)^2 & \text{if } c = 10 \text{ and } (\Delta_{\text{cov}} > 1 \text{ or } \Delta_{\text{tail}} > 2) \\ 2.0 \cdot (H - L) \cdot \left(\frac{\Delta_c}{5}\right)^2 & \text{otherwise} \end{cases}$$

Rationale: Confidence calibration rewards honest self-assessment. The ideal confidence mapping establishes that perfect 10% outside coverage merits maximum confidence (10), while progressively poorer coverage should correspond to lower claimed confidence. The special case for $c = 10$ implements strict requirements: claiming certainty of perfection requires both accurate total coverage ($\Delta_{\text{cov}} < 1\%$) and balanced tails ($\Delta_{\text{tail}} < 2\%$), enforced through a 3.0× penalty multiplier.

Properties: - Maps coverage quality to expected confidence level - Quadratic penalty for calibration mismatch - Enhanced scrutiny for maximum confidence claims - Proportional to interval width

3.5 Balance Penalty

Definition:

$$P_{\text{balance}} = 1.0 \cdot (H - L) \cdot \left(\frac{\Delta_{\text{tail}}}{10} \right)^2$$

where $\Delta_{\text{tail}} = |P_{\text{below}} - 5| + |P_{\text{above}} - 5|$.

Rationale: A well-specified 90% CI should exhibit approximate symmetry with 5% in each tail. This component explicitly penalizes imbalanced intervals, preventing gaming strategies where, for example, a very tight lower bound compensates for an excessively loose upper bound to achieve 10% total outside. The penalty is applied universally, independent of confidence level.

Properties: - Quadratic in tail imbalance - Enforces symmetric coverage - Independent of confidence calibration - Coefficient 1.0 provides moderate weight

3.6 Complete Scoring Function

3.7 Mathematical Specification

The complete scoring function for interval $[L, H]$ with confidence c and dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ is:

$$\boxed{S(L, H, c \mid \mathcal{D}) = P_{\text{precision}} + P_{\text{miss}} + P_{\text{over}} + P_{\text{calib}} + P_{\text{balance}}}$$

where all penalty components are defined as above. The score is rounded to the nearest integer for reporting.

3.8 Optimal Strategy

The scoring function incentivizes the following strategy:

1. **Accuracy:** Match estimated bounds to actual sample percentiles (p5 and p95) constraints
2. **Coverage:** Achieve approximately 90% empirical coverage
3. **Calibration:** Match claimed confidence to actual interval quality
4. **Balance:** Distribute coverage symmetry with 5% in each tail
5. **Honesty:** Report confidence levels that reflect genuine uncertainty

Deviations from any dimension incur quadratic penalties, with weights tuned to prevent dominance by any single component.

4 Illustrative Examples

4.1 Example: Perfect Interval

Specifications:

- Sample percentiles: $p_5 = 1080$ seconds (18 min), $p_{95} = 1680$ seconds (28 min)
- Estimated interval: $[L, H] = [1080, 1680]$ (exactly matching sample percentiles)
- Empirical coverage: 90.2% (4.8% below, 5.0% above)
- Confidence: $c = 10$

Penalty Calculation:

- $P_{\text{precision}} = (1080 - 1080)^2 + (1680 - 1680)^2 = 0$ (perfect match!)
- $P_{\text{miss}} \approx 150$ (small residual from 9.8% outside)
- $P_{\text{over}} = 0$ (coverage $< 95\%$)
- $P_{\text{calib}} \approx 0$ ($\Delta_{\text{cov}} = 0.2 < 1$, $\Delta_{\text{tail}} = 0.2 < 2$)

- $P_{\text{balance}} \approx 0.07$ ($\Delta_{\text{tail}} = 0.2$)

Total Score: $S \approx 150$

Interpretation: Near-optimal performance with minimal penalties. The precision penalty is zero because the estimates perfectly match the sample percentiles.

4.2 Example: Overconfident Estimate

Specifications:

- Sample percentiles: $p_5 = 1080$ seconds, $p_{95} = 1680$ seconds
- Estimated interval: $[L, H] = [1200, 1800]$ (120 sec high on each bound)
- Empirical coverage: 85% (8% below, 7% above)
- Confidence: $c = 10$

Penalty Calculation:

- $P_{\text{precision}} = (1080 - 1200)^2 + (1680 - 1800)^2 = 14400$ (poor match!)
- $P_{\text{miss}} \approx 400$ (15% outside)
- $P_{\text{over}} = 0$
- $\Delta_{\text{cov}} = |15 - 10| = 5 > 1$
- $\Delta_{\text{tail}} = |8 - 5| + |7 - 5| = 5 > 2$
- Interval width for other penalties: $1800 - 1200 = 600$ seconds
- $P_{\text{calib}} = 3.0 \cdot 600 \cdot \left(\frac{5+5}{5}\right)^2 = 7200$ (heavy penalty)
- $P_{\text{balance}} = 1.0 \cdot 600 \cdot \left(\frac{5}{10}\right)^2 = 150$

Total Score: $S \approx 36550$

Interpretation: Large precision penalty for poor boundary estimation, plus severe penalty for claiming perfection (confidence = 10) with poor actual performance.

4.3 Example: Unbalanced Interval

Specifications:

- Sample percentiles: $p_5 = 1080$ seconds, $p_{95} = 1680$ seconds
- Estimated interval: $[L, H] = [1050, 1730]$ (30 sec low, 50 sec high)
- Empirical coverage: 90% (2% below, 8% above)
- Confidence: $c = 9$

Penalty Calculation:

- $P_{\text{precision}} = (1050 - 1080)^2 + (1730 - 1680)^2 = 900 + 2500 = 3400$
- $P_{\text{miss}} \approx 200$ - $P_{\text{over}} = 0$
- $P_{\text{miss}} \approx 200$ - $P_{\text{over}} = 0$
- $\Delta_{\text{cov}} = 0$ (10% outside) - $c_{\text{ideal}} = 10$, $\Delta_c = 1$
- Interval width: $1730 - 1050 = 680$ seconds
- $P_{\text{calib}} = 2.0 \cdot 680 \cdot \left(\frac{1}{5}\right)^2 = 54$
- $\Delta_{\text{tail}} = |2 - 5| + |8 - 5| = 6$

- $P_{\text{balance}} = 1.0 \cdot 680 \cdot \left(\frac{6}{10}\right)^2 = 245$

Total Score: $S \approx 3899$

Interpretation: Moderate penalty for tail imbalance and imperfect boundary estimation despite correct total coverage.

4.4 Example: Honest Uncertainty

Specifications:

- Sample percentiles: $p_5 = 1080$ seconds, $p_{95} = 1680$ seconds
- Estimated interval: $[L, H] = [900, 1900]$ (180 sec low, 220 sec high - very conservative)
- Empirical coverage: 75% (12% below, 13% above)
- Confidence: $c = 7$

Penalty Calculation:

- $P_{\text{precision}} = (900 - 1080)^2 + (1900 - 1680)^2 = 32400 + 48400 = 80800$
- $P_{\text{miss}} \approx 600$ - $P_{\text{over}} = 0$
- $\Delta_{\text{cov}} = |25 - 10| = 15$
- $c_{\text{ideal}} = 10 - 15/10 = 8.5$, $\Delta_c = 1.5$
- Interval width: $1900 - 900 = 1000$ seconds
- $P_{\text{calib}} = 2.0 \cdot 1000 \cdot \left(\frac{1.5}{5}\right)^2 = 180$
- $\Delta_{\text{tail}} = |12 - 5| + |13 - 5| = 15$
- $P_{\text{balance}} = 1.0 \cdot 1000 \cdot \left(\frac{15}{10}\right)^2 = 2250$

Total Score: $S \approx 83830$

Interpretation: Very poor boundary estimation (large precision penalty) combined with poor interval quality, though relatively small calibration penalty due to appropriately modest confidence claim.

5 Strengths and Limitations

5.1 Strengths

1. **Multi-dimensional Assessment:** The scoring system simultaneously evaluates estimation accuracy, coverage, calibration, and balance, preventing optimization of any single dimension at the expense of others.
2. **Gaming Resistance:** Specific penalty components (overcoverage, balance) explicitly counter known gaming strategies such as arbitrarily wide intervals or asymmetric bounds.
3. **Theoretical Grounding:** The quadratic penalty structure inherits desirable properties from Brier scoring, including proper scoring rule characteristics that encourage honest reporting.
4. **Interpretable Components:** Each penalty term corresponds to an intuitive quality dimension with clear physical interpretation.
5. **Calibration Incentive:** The confidence calibration component explicitly rewards self-awareness and penalizes overconfidence, promoting metacognitive skill development.
6. **Robustness:** The 10% trimming in miss penalty calculation provides resilience to extreme outliers while maintaining sensitivity to systematic coverage failures.

5.2 Limitations

1. **Parameter Sensitivity:** The relative weights of penalty components (0.5, 1.0, 2.0, 3.0) were chosen heuristically. Alternative weightings may produce different optimal strategies, and formal optimization of these hyperparameters remains an open question.
2. **Distributional Assumptions:** The 5%-5% tail balance target implicitly assumes approximate symmetry in the underlying distribution. For highly skewed commute time distributions, this requirement may be overly restrictive.
3. **Sample Size Dependence:** With small sample sizes ($n \approx 20 - 30$), sample percentiles exhibit estimation variance. The scoring function does not explicitly account for this uncertainty in the target percentiles.
4. **Non-Convexity:** The piecewise nature of some penalty functions (e.g., overcoverage threshold) introduces discontinuities that may complicate optimization and interpretation near boundary regions.
5. **Scale Dependence:** Penalty magnitudes scale with interval width, which inherently depends on the time scale of the underlying data. Cross-dataset comparisons require careful normalization.
6. **Confidence Discretization:** Restricting confidence to integer values in $[5, 10]$ provides limited resolution for fine-grained calibration assessment. A continuous scale might better capture nuanced uncertainty. This needs to be balanced against the poor ability of most individuals to reliably distinguish small differences in confidence.
7. **Independence Assumption:** The additive penalty structure assumes independence among quality dimensions. In practice, certain trade-offs are inevitable, and the optimal balance depends on the specific weight configuration.

5.3 Relation to Traditional Brier Scoring

Both the precision penalty ($P_{\text{precision}}$) and miss penalty (P_{miss}) implement Brier's mean squared error principle:

$$P_{\text{precision}} = (L - p_5)^2 + (H - p_{95})^2$$

$$P_{\text{miss}} = \frac{1}{n} \sum_{m \in \mathcal{M}'} m = \frac{1}{n} \sum_{m \in \mathcal{M}'} (\text{distance})^2$$

This parallels the Brier score's $(f - o)^2$ formulation, where forecast error is penalized quadratically. However, our extension differs in several key aspects:

1. **Continuous vs. Binary:** We evaluate continuous interval bounds rather than binary event probabilities.
2. **Multiple Dimensions:** Traditional Brier scores measure a single forecast-outcome discrepancy, whereas our system assesses five distinct quality aspects.
3. **Subjective Calibration:** The confidence calibration component introduces a meta-level assessment of forecaster self-awareness, absent in standard Brier scoring.
4. **Structural Penalties:** Components like balance and overcoverage penalties enforce structural properties beyond simple accuracy.

The quadratic penalty structure remains central: larger deviations are disproportionately costly, encouraging forecasters to minimize expected squared error across all dimensions. This maintains the proper scoring rule philosophy that honest, well-calibrated estimates should minimize expected penalty.

6 Implementation Considerations

6.1 Computational Complexity

The scoring function requires $O(n \log n)$ time due to sorting operations in miss penalty calculation. For typical commute datasets ($n \approx 50 - 200$), this is negligible. Space complexity is $O(n)$ for storing miss values.

6.2 Numerical Stability

All penalty calculations use floating-point arithmetic. Care must be taken with: - Division by zero when $n = 0$ (handled by minimum record requirement $n \geq 20$) - Overflow in squared distance calculations for extreme outliers (mitigated by 10% trimming)

6.3 Verification and Reproducibility

A cryptographic checksum mechanism is provided to verify score integrity:

$$\text{checksum} = \text{Hash}(L, H, S, c)$$

where the hash function combines rounded parameter values into a deterministic 32-bit signature. This enables external verification via Excel VBA or other independent implementations, ensuring audit trail integrity for competitive challenges.

7 Conclusion

We have presented a comprehensive scoring methodology for confidence interval estimation that extends Brier scoring principles to multi-dimensional interval quality assessment. The system effectively balances competing objectives through weighted quadratic penalties while explicitly countering gaming strategies. Through careful calibration of penalty weights and special handling of extreme confidence claims, the scoring function encourages honest, well-calibrated predictions with balanced coverage.

The methodology demonstrates that proper scoring rules can be extended beyond binary events to complex estimation tasks requiring simultaneous optimization across multiple quality dimensions. Future work might explore adaptive penalty weights based on sample size, explicit uncertainty quantification in score reporting, and extensions to non-symmetric target distributions.

8 References

1. Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability." *Monthly Weather Review*, 78(1), 1-3.
2. Gneiting, T., & Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477), 359-378.
3. Winkler, R. L. (1996). "Scoring rules and the evaluation of probabilities." *Test*, 5(1), 1-60.
4. Merkle, E. C., & Steyvers, M. (2013). "Choosing a strictly proper scoring rule." *Decision Analysis*, 10(4), 292-304.

Document Version Control: - v0.1 (2025-11-04): Initial specification and mathematical formulation - Pre-Publication. Awaiting review.