Journal of Machine Learning Research 23 (2022) 1-54 Submitted 6/22; Revised 10/22; Published 10/22

Metrics of Calibration for Probabilistic Predictions

Imanol Arrieta-Ibarra IMANOLAI@META.COM
Paman Gujral PGUJRAL@META.COM
Jonathan Tannen JONATHAN.TANNEN@GMAIL.COM
Mark Tygert* TYGERT@META.COM
Cherie Xu CHERIEXU@META.COM
Meta, 1 Facebook Way, Menlo Park, CA 94025, USA
**Please send correspondence to* MARK@TYGERT.COM

Editor: Sayan Mukherjee

Abstract

Many predictions are probabilistic in nature; for example, a prediction could be for precipitation tomorrow, but with only a 30% chance. Given such probabilistic predictions together with the actual outcomes, "reliability diagrams" (also known as "calibration plots") help detect and diagnose statistically significant discrepancies—so-called "miscalibration"—between the predictions and the outcomes. The canonical reliability diagrams are based on histogramming the observed and expected values of the predictions; replacing the hard histogram binning with soft kernel density estimation using smooth convolutional kernels is another common practice. But, which widths of bins or kernels are best? Plots of the cumulative differences between the observed and expected values largely avoid this question, by displaying miscalibration directly as the slopes of secant lines for the graphs. Slope is easy to perceive with quantitative precision, even when the constant offsets of the secant lines are irrelevant; there is no need to bin or perform kernel density estimation.

The existing standard metrics of miscalibration each summarize a reliability diagram into a single scalar statistic. The cumulative plots naturally lead to scalar metrics for the deviation of the graph of cumulative differences away from zero; good calibration corresponds to a horizontal, flat graph which deviates little from zero. The cumulative approach is currently unconventional, yet offers many favorable statistical properties, guaranteed via mathematical theory backed by rigorous proofs and illustrative numerical examples. In particular, metrics based on binning or kernel density estimation unavoidably must trade-off statistical confidence for the ability to resolve variations as a function of the predicted probability or vice versa. Widening the bins or kernels averages away random noise while giving up some resolving power. Narrowing the bins or kernels enhances resolving power while not averaging away as much noise. The cumulative methods do not impose such an explicit trade-off. Considering these results, practitioners probably should adopt the cumulative approach as a standard for best practices.

Keywords: reliability diagram, calibration plot, cumulative differences, Kolmogorov-Smirnov, Kuiper

1. Introduction

Given 100 independent observations of outcomes ("success" or "failure") of Bernoulli trials that are forecast to have an 80% chance of success, the forecasts are perfectly *calibrated* when 80 of the observations report success. More generally, given some number, say n, of independent observations of outcomes of Bernoulli trials that are forecast to have a probability S of success, the predictions are perfectly calibrated when nS of the observations report success. Needless to say, the actual number of observations of success is likely to vary around nS randomly, so in practice we test not whether nS is exactly equal to the observed number of successes, but rather whether the difference

between nS and the observed number of successes is statistically significant. Such significance tests can be found in any standard textbook on statistics.

The present paper considers the following more general setting: Suppose we have n observations R_1, R_2, ..., R_n of the outcomes of independent Bernoulli trials with corresponding predicted probabilities of success, say S_1, S_2, ..., S_n. For instance, each S_k could be a classifier's probabilistic score and the corresponding R_k could be the indicator of correct classification, with $R_k = 1$ when the classification is correct and $R_k = 0$ when the classification is incorrect (so R_k could also be regarded as a class label, where class 1 corresponds to "the classifier succeeded" and class 0 corresponds to "the classifier erred"). We would then want to test the hypothesis that the response R_k is distributed as a Bernoulli variate with expected value S_k for all $k = 1, 2, \ldots, n$, which is the same null hypothesis considered in the previous paragraph when $S_1 = S_2 = \cdots = S_n = S$. The remainder of this article simplifies the analysis by reordering the samples (preserving the pairing of R_k with S_k for every k) such that $S_1 \le S_2 \le \cdots \le S_n$, with any ties ordered randomly, perturbing so that $S_1 < S_2 < \cdots < S_n$.

One method for measuring calibration is to histogram the responses R_1, R_2, ..., R_n as a function of the scores S_1, S_2, ..., S_n; this involves partitioning the scores into m sets known as "bins" (or "buckets") and calculating both the average score and the average response for the scores and corresponding responses falling in each bin. The conventional scalar-valued metrics for calibration are known as "empirical calibration errors" or "estimated calibration errors" (ECEs). "ECE^1" denotes the sum over every histogram bin of the absolute value of the difference between the average response and the average score, each weighted by the width of the bin. Thus, the ECE^1 estimates the area between the observed responses and ideal calibration. "ECE^2" denotes the sum that is the same aside from squaring every difference prior to summation.

Another method for measuring calibration is to graph the cumulative differences between the responses and the scores. The expected slope of any secant line connecting two points on the graph is equal to the average miscalibration over the scores between those points. In the case of perfect calibration (for which each response is equal to the corresponding score), the resulting graph is a horizontal, flat line at zero. Both the maximum deviation of the graph from zero and the range (maximum minus minimum) of the graph measure the deviation of the graph from the horizontal, flat ideal of perfect calibration. We call these metrics "empirical cumulative calibration errors" or "estimated cumulative calibration errors" (ECCEs), with the maximum absolute deviation being denoted "ECCE-MAD" and the range of deviations being denoted "ECCE-R". Our earlier publications referred to "ECCE-MAD" as the "Kolmogorov-Smirnov" statistic and to "ECCE-R" as the "Kuiper" statistic.

The present paper follows up and elaborates on problems highlighted earlier by Gupta et al. (2020), Roelofs et al. (2020), Tygert (2021a), and Tygert (2021b); they point out that the classical empirical calibration errors based on binning vary significantly based on the choice of bins. The choice of bins is fairly arbitrary and enables the analyst to fudge results (whether purposefully or unintentionally). Having to make such a critical yet arbitrary choice is especially fraught when dealing with laws and regulators seeking a universal standard for compliance. Also concerning is the bias observed by Roelofs et al. (2020) in the estimates given by empirical calibration errors based on binning. The present work proves that the ECCEs have no bias like the ECEs'.

The results of the present article are all implicit in those of Bröcker (2008), Gupta et al. (2020), Kumar et al. (2019), Nixon et al. (2019), Roelofs et al. (2020), Simonoff and Udina (1997), Srihera and Stute (2010), Stute (1997), and Vaicenavicius et al. (2019). The purpose of the present paper is to provide a simple, rigorous exposition of what might be viewed as a unifying thread throughout the other works. In particular, this article directly compares the ECEs to the ECCEs, more extensively than prior work has.

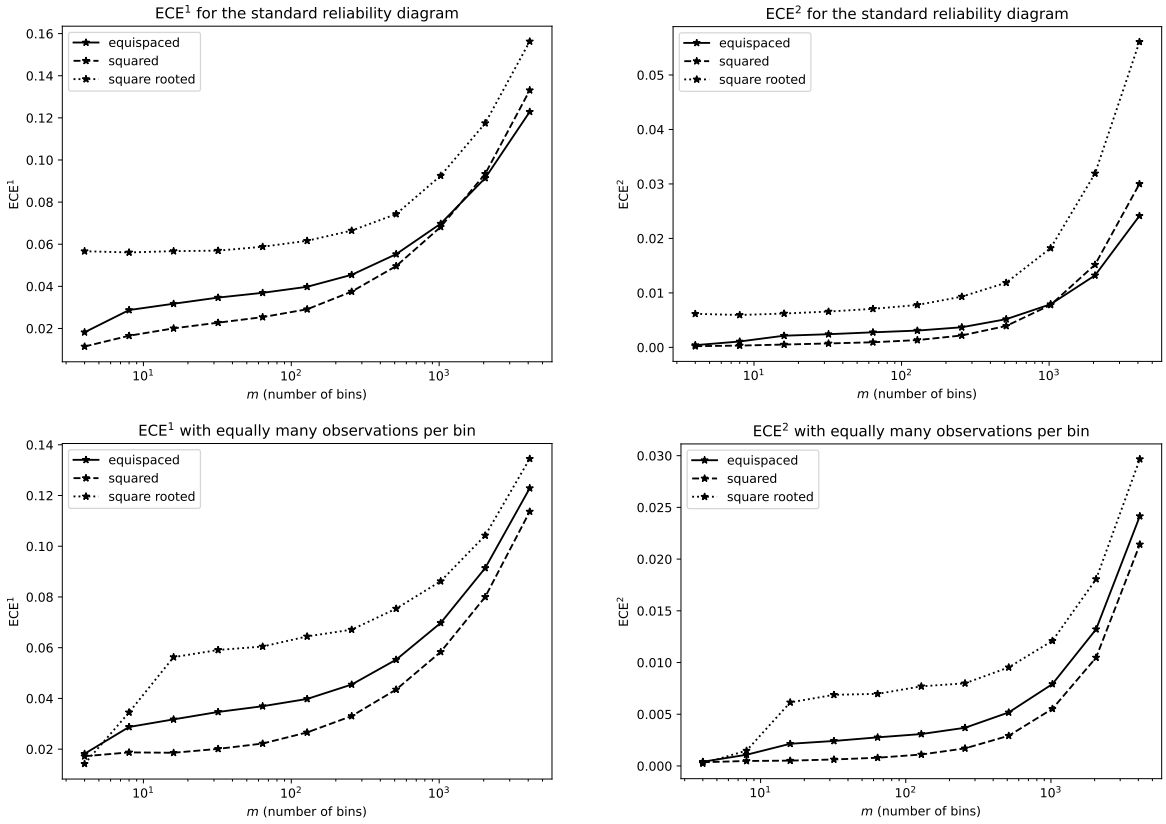

Figure 1: Empirical calibration errors for the data set of Section 3.1 with sample size $n = 32{,}768$; the scores are equispaced, squared after initially being equispaced, or square rooted after initially being equispaced, as indicated in the legends.

The empirical calibration errors based on binning intrinsically trade-off resolution for statistical confidence or vice versa. Widening the bins averages away more noise in the estimates, while sacrificing some of the power to resolve variations as a function of score. Narrowing the bins resolves finer variations as a function of score, while not averaging away as much noise in the estimates. In contrast, the empirical *cumulative* calibration errors exhibit no such explicit trade-off, with no parameters to adjust. The empirical cumulative calibration errors are fully non-parametric and uniquely, fully specified, statistically powerful and reliable.

The present paper directly treats unweighted samples. To be sure, the results of the present article generalize to the case of weighted samples. In that case, each observation comes with a positive real number that indicates how heavily to weight the observation when combining it with the other observations in expected values. However, weighted sampling introduces additional complications that distract from the comparison between the standard binned metrics and the cumulative metrics, so the present article focuses on the case of unweighted sampling. (Of course, unweighted sampling is equivalent to uniform weighting, in which all weights are equal.) Extensive graphical comparisons for the case of weighted sampling are available from Tygert (2021a) and Tygert (2021b).

The paper is organized as follows: the next section, Section 2, details the methodologies and proves theorems about their performance. Then, Section 3 illustrates the methodologies and the-

ory of Section 2 via several examples, using both synthetic and measured data sets.[1] Finally, Section 4 reviews the results, drawing conclusions. Appendix A gives detailed, rigorous proofs, and Appendix B supplements the plots of Section 3 with additional plots.

2. Methods

This section gives theorems characterizing advantages of cumulative metrics over the standard binned metrics. A full exposition requires the detailed, rigorous proofs discussed in this section; however, the high-level strategy of all the derivations is quite simple and straightforward, and can be summarized as follows:[2]

First, Section 2.1 defines both standard binned metrics and cumulative metrics for assessing deviation from perfect calibration. For a *perfectly* calibrated underlying distribution, both the ideal, underlying calibration error and the ideal, underlying cumulative calibration error are equal to 0, whereas for an *imperfectly* calibrated underlying distribution, both are greater than 0.

Next, Section 2.2 proves that, for a *perfectly* calibrated underlying distribution, as the sample size increases without bound the cumulative metrics converge to 0 if the empirical cumulative distribution function of the scores converges uniformly to a continuous cumulative distribution function. Section 2.2 also proves that the expected values of the standard binned metrics stay bounded away from 0 if the number of draws per bin remains bounded on some fixed range of scores, as the maximum bin width becomes arbitrarily small.

Then, Section 2.3 proves that, for an *imperfectly* calibrated underlying distribution, as the sample size becomes arbitrarily large the expected values of the cumulative metrics stay bounded away from 0. (The proof is under the assumption that the empirical cumulative distribution function of the scores converges uniformly to a cumulative distribution function.) Section 2.3 also proves that the expected values of the standard binned metrics stay bounded away from 0 if the maximum bin width becomes arbitrarily small. This shows that the cumulative metrics can distinguish any imperfectly calibrated underlying distribution from perfect calibration, given enough observed scores and associated responses. This also shows that there exist imperfectly calibrated distributions which the standard binned metrics cannot distinguish from perfect calibration as the maximum width of a bin approaches 0, if the number of draws per bin remains bounded on some fixed range of scores.

Section 2.4 then derives the consequences, namely, that the standard binned metrics are intrinsically subject to an unavoidable trade-off, requiring observations whose density is infinitely higher than the density that the cumulative statistics need, in the limit required for asymptotic consistency. Moreover, the trade-off is even more unwieldy when the number of observations is limited: the standard binned metrics vary significantly with the (rather arbitrary) choice of bins, whereas the cumulative metrics work uniformly well without any tuning. In fact, the cumulative metrics require no tuning at all—the cumulative metrics have no tuning parameters whatsoever; fudging their results is simply impossible.

Finally, Section 2.5 summarizes the usual motivations for the constructions of these particular metrics, reviewing the associated graphical methods.

Most of the rest of the present section, Sections 2.1–2.4, now provides rigorous details of this strategy. The notational conventions used throughout the present paper adhere to those set in Section 2.1. Section 2.5 briefly reviews the principal motivations for considering the particular metrics studied here.

1. Permissively licensed open-source software that automatically reproduces all figures and statistics reported below is available at `https://github.com/facebookresearch/ecevecce`
2. All sections of the present paper use the same notation that Section 2.1 introduces.

2.1 Definitions

This subsection defines in detail the metrics analyzed below. Section 2.1.1 defines the standard binned statistics, while Section 2.1.2 defines the cumulative statistics. Graphical methods summarized in Section 2.5 motivate all these definitions, though the present subsection omits discussion of the motivation in order to keep the exposition concise and easily digestible for those already familiar with the motivations.

2.1.1 Binned

The observations come as pairs of scores and responses. Each score is a real number in the unit interval $[0, 1]$; each response is a random variable whose value is either 0 or 1. The positive integer m will denote the number of bins; bin j contains the scores S_j^k for $k = 1, 2, \ldots, n_j$, for $j = 1, 2, \ldots, m$. Each score S_j^k comes with a response R_j^k; the responses are independent and, under the null hypothesis of perfect calibration, R_j^k is a Bernoulli variate whose expected value is S_j^k (the variance is then $S_j^k(1 - S_j^k)$). We order the scores such that $S_i^k < S_j^\ell$ whenever $i < j$, and $S_i^k < S_j^\ell$ when $i = j$ and $k < \ell$. For notational convenience, we define $S_0^1 = 0$ and $S_{m+1}^1 = 1$. We denote by n the total number of observations, that is, $n = n_1 + n_2 + \cdots + n_m$. *We suppress the sequence index n for S_j^k, R_j^k, n_j, and m; in principle the full notation would be $S_j^k(n)$, $R_j^k(n)$, $n_j(n)$, and $m(n)$, where $n = 1, 2, 3, \ldots$*; below, "convergence" refers to convergence with respect to the sample size n increasing without bound. The scores for each sample size n are assumed to be distinct. The average score in bin j is

$$\tilde{S}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} S_j^k, \tag{1}$$

while the average response in bin j is

$$\tilde{R}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} R_j^k. \tag{2}$$

The mean-square empirical calibration error (ECE^2) is the Riemann sum

$$\mathrm{ECE}^2 = \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\tilde{R}_j - \tilde{S}_j\right)^2 = \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j}\right)^2; \tag{3}$$

analogously, the l^1 empirical calibration error (ECE^1) is the Riemann sum

$$\mathrm{ECE}^1 = \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left|\tilde{R}_j - \tilde{S}_j\right| = \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left|\sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j}\right|. \tag{4}$$

2.1.2 Cumulative

We now define $S_1, S_2, \ldots, S_n$ to be all $n = \sum_{j=1}^{m} n_j$ of the scores S_j^k, sorted in ascending order. Sorting in ascending order means that $S_1 < S_2 < \cdots < S_n$ (recall from Section 1 that the scores are distinct). Furthermore, if S_ℓ is the score equal to S_j^k, then we set R_ℓ to be equal to the corresponding response, R_j^k. For notational convenience, we also define $S_0 = 0$ and $S_{n+1} = 1$. *As in the binned case, we suppress the sequence index n for S_ℓ and R_ℓ; in principle the full notation would be $S_\ell(n)$ and $R_\ell(n)$, where $n = 1, 2, 3, \ldots$*; as mentioned earlier, "convergence" will refer to convergence

with respect to the index n increasing without bound. Again, the scores for each sample size n are assumed to be distinct. The cumulative differences are

$$C_k = \frac{1}{n} \sum_{j=1}^{k} (R_j - S_j) \tag{5}$$

for $k = 1, 2, \ldots, n$. The maximum absolute deviation of the empirical cumulative calibration error (ECCE-MAD) is

$$\text{ECCE-MAD} = \max_{1 \le k \le n} |C_k| \tag{6}$$

and the range of the empirical cumulative calibration error (ECCE-R) is

$$\text{ECCE-R} = \max_{0 \le k \le n} C_k - \min_{0 \le k \le n} C_k, \tag{7}$$

where C_k is defined in (5) and $C_0 = 0$. Another term for "ECCE-MAD" is the "Kolmogorov-Smirnov metric," and another term for "ECCE-R" is the "Kuiper metric"—Kolmogorov (1933), Smirnov (1939), and Kuiper (1962) introduced these statistics in order to solve a similar problem, namely, determining the statistical significance of observed differences in empirical probability distributions.

The absolute value of the total miscalibration $\sum_{j \in I} (R_j - S_j)/n$ over any interval I of indices is less than or equal to the ECCE-R; indeed, the ECCE-R is the maximum of the absolute value of the total miscalibration over any interval of indices: $\text{ECCE-R} = \max_I |\sum_{j \in I} (R_j - S_j)/n|$.

2.2 Perfectly calibrated responses

This subsection analyzes the expected values of the metrics when the responses arise from a perfectly calibrated distribution, that is, under the assumption of the null hypothesis of perfect calibration. The principal results of this subsection are Corollaries 2 and 4 for the ECE and Corollary 7 for the ECCE.

2.2.1 Binned

The following theorem provides the lower limit of the ECE^2, yielding the very useful Corollary 2. Appendix A details the proof of the theorem.

Theorem 1 *Assume the null hypothesis that the expected value of the response R_j^k is equal to the corresponding score S_j^k for all $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n_j$. Suppose also that $\max_{0 \le j \le m} (S_{j+1}^1 - S_j^1)$ converges to 0 as the sample size n increases without bound, and that $\nu = \nu_n$ is the step function starting from $\nu(0) = \nu(S_0^1) = n_1$ and jumping to $\nu(S_j^k) = n_j$ for all $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n_j$. Then, as n becomes arbitrarily large the lower limit of the expected value of the ECE^2 defined in (3) converges to*

$$\liminf_{n \to \infty} \int_0^1 \frac{s(1-s)}{\nu(s)} \, ds, \tag{8}$$

where lim inf *denotes the lower limit and the sequence index n of the function $\nu = \nu_n$ is suppressed in the notation.*

The main consequence of this theorem is the following, stating that the ECE^2 hits a "noise floor."

Corollary 2 *If $\nu = \nu_n$ is bounded from above on some interval, independent of the sample size n, and is the step function starting from $\nu(0) = \nu(S_0^1) = n_1$ and jumping to $\nu(S_j^k) = n_j$ for all $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n_j$, then the expected value of the ECE^2 defined in (3) is greater than a fixed strictly positive real number for all sufficiently large n, assuming the null hypothesis of perfect calibration and that the bin width $(S_{j+1}^1 - S_j^1)$ converges uniformly to 0 (uniformly over $j = 0, 1, \ldots, m(n)$).*

A similar corollary holds for the ECE^1 defined in (4), due to the following theorem. Appendix A details the proof of the theorem.

Theorem 3 *The ECE^1 defined in (4) is greater than or equal to the ECE^2 defined in (3).*

Combining Theorem 3 and Corollary 2 yields the following, stating that the ECE^1 hits a "noise floor."

Corollary 4 *If $\nu = \nu_n$ is bounded from above on some interval, independent of the sample size n, and is the step function starting from $\nu(0) = \nu(S_0^1) = n_1$ and jumping to $\nu(S_j^k) = n_j$ for all $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n_j$, then the expected value of the ECE^1 defined in (4) is greater than a fixed strictly positive real number for all sufficiently large n, assuming the null hypothesis of perfect calibration and that the bin width $(S_{j+1}^1 - S_j^1)$ converges uniformly to 0 (uniformly over $j = 0, 1, \ldots, m(n)$).*

2.2.2 Cumulative

The independence of the responses yields the following theorem, which yields Corollary 7 when combined with Theorem 6.

Theorem 5 *Assume the null hypothesis that the expected value of the response R_k is equal to the corresponding score S_k for all $k = 1, 2, \ldots, n$. Then, the variance of C_k defined in (5) is*

$$(\sigma_k)^2 = \frac{1}{n^2} \sum_{j=1}^{k} S_j(1 - S_j) \leq \frac{k}{4n^2} \leq \frac{1}{4n} \tag{9}$$

for $k = 1, 2, \ldots, n$.

The following theorem summarizes Sections 2.3 and 3 of Tygert (2022).

Theorem 6 *Assume the null hypothesis that the expected value of the response R_k is equal to the corresponding score S_k for all $k = 1, 2, \ldots, n$. Suppose again that the scores $S_1, S_2, \ldots, S_n$ are all distinct for each sample size n, and also that $\max_{1 \leq k \leq n} S_k(1 - S_k) / \sum_{j=1}^{n} S_j(1 - S_j)$ converges to 0 as n increases without bound.*

Then, as n becomes arbitrarily large the ECCE-MAD divided by σ_n converges in distribution to the maximum of the absolute value of the standard Brownian motion over the unit interval $[0, 1]$. The ECCE-MAD is defined in (6) and σ_n is defined in (9).

Moreover, as n increases without bound the ECCE-R divided by σ_n converges in distribution to the range of the standard Brownian motion over the unit interval $[0, 1]$. The ECCE-R is defined in (7) and σ_n is defined in (9).

The expected value of the maximum of the absolute value of the standard Brownian motion over $[0, 1]$ is $\sqrt{\pi/2} \approx 1.25$, and the expected value of the range of the standard Brownian motion over $[0, 1]$ is $2\sqrt{2/\pi} \approx 1.60$.

Combining Theorems 5 and 6 yields the following.

Corollary 7 *As n becomes arbitrarily large both the ECCE-MAD defined in (6) and the ECCE-R defined in (7) converge to 0, assuming both the null hypothesis of perfect calibration and that the scores S_1, S_2, ..., S_n are all distinct for each sample size n, as well as that $\max_{1\le k\le n} S_k(1-S_k)/\sum_{j=1}^n S_j(1-S_j)$ converges to 0 as n increases without bound.*

2.3 Imperfectly calibrated responses

This subsection analyzes the expected values of the metrics when the responses arise from an imperfectly calibrated distribution, that is, under the assumption of an "alternative" hypothesis that differs nontrivially from the null hypothesis of perfect calibration. The principal results of this subsection are Corollaries 9 and 10 for the ECE and Theorem 11 for the ECCE.

2.3.1 Binned

Theorem 1 above is similar to the following theorem for the special case in which the function r in this coming theorem satisfies $r(s) = s$. Appendix A details the proof of the theorem; the proof is similar to that of Theorem 1, albeit significantly longer.

Theorem 8 *Suppose that $r : [0,1] \to [0,1]$ is any piecewise continuous function that is independent of the sample size n. Suppose also that the response R_j^k is drawn from the Bernoulli distribution whose expected value is $r(S_j^k)$ for all $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n_j$. Suppose finally that $\max_{0\le j\le m}(S_{j+1}^1 - S_j^1)$ converges to 0 as n increases without bound, and that $\nu = \nu_n$ is the step function starting from $\nu(0) = \nu(S_0^1) = n_1$ and jumping to $\nu(S_j^k) = n_j$ for all $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n_j$. Then, as n becomes arbitrarily large the lower limit of the expected value of the ECE^2 defined in (3) converges to*

$$\liminf_{n\to\infty} \int_0^1 \left((r(s) - s)^2 + \frac{r(s)\,(1 - r(s))}{\nu(s)} \right) ds, \tag{10}$$

where lim inf *denotes the lower limit and the sequence index n of the function $\nu = \nu_n$ is suppressed in the notation.*

As with Theorem 1 and Corollary 2, the main consequence of Theorem 8 is the following.

Corollary 9 *Suppose that $r : [0,1] \to [0,1]$ is any piecewise continuous function that is independent of the sample size n. Suppose also that the response R_j^k is drawn from the Bernoulli distribution whose expected value is $r(S_j^k)$ for all $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n_j$. If r is also imperfectly calibrated, that is, $\int_0^1 (r(s)-s)^2\,ds > 0$, then the expected value of the ECE^2 defined in (3) is greater than a fixed strictly positive real number for all sufficiently large n, assuming that the bin width $(S_{j+1}^1 - S_j^1)$ converges uniformly to 0 (uniformly over $j = 0, 1, \ldots, m(n)$).*

Combining Theorem 3 and Corollary 9 yields the following similar result for the ECE^1.

Corollary 10 *Suppose that $r : [0,1] \to [0,1]$ is any piecewise continuous function that is independent of the sample size n. Suppose also that the response R_j^k is drawn from the Bernoulli distribution whose expected value is $r(S_j^k)$ for all $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n_j$. If r is also imperfectly calibrated, that is, $\int_0^1 (r(s)-s)^2\,ds > 0$, then the expected value of the ECE^1 defined in (4) is greater than a fixed strictly positive real number for all sufficiently large n, assuming that the bin width $(S_{j+1}^1 - S_j^1)$ converges uniformly to 0 (uniformly over $j = 0, 1, \ldots, m(n)$).*

2.3.2 Cumulative

The following theorem is an analogue for the ECCE of Corollaries 9 and 10. Appendix A details the proof of the theorem.

Theorem 11 *Suppose that the empirical cumulative distribution function of the scores S_1, S_2, ..., S_n converges uniformly to some cumulative distribution function F as the sample size n increases without bound. Suppose further that $r : [0,1] \to [0,1]$ is any function that is independent of the sample size n, is Riemann-Stieltjes integrable with respect to F, and is imperfectly calibrated, that is, $\int_0^1 |r(s) - s| \, dF(s) > 0$. Suppose also that the response R_k is drawn from the Bernoulli distribution whose expected value is $r(S_k)$ for all $k = 1, 2, \ldots, n$, and that $r(S_0) = S_0$ and $r(S_{n+1}) = S_{n+1}$ (where $S_0 = 0$ and $S_{n+1} = 1$). Then, the expected values of the ECCE-MAD defined in (6) and of the ECCE-R defined in (7) stay bounded away from 0 for all sufficiently large n.*

2.4 Consequences

This subsection puts together the main results of this section.

For a *perfectly* calibrated underlying distribution, both the ideal, underlying calibration error and the ideal, underlying cumulative calibration error are equal to 0, whereas for an *imperfectly* calibrated underlying distribution, both are greater than 0. For a *perfectly* calibrated underlying distribution, Corollaries 2 and 4 show that the expected values of the ECE^1 and of the ECE^2 stay bounded away from 0 if the number of draws per bin remains bounded on some fixed range of scores, as the maximum bin width becomes arbitrarily small, while Corollary 7 shows that as the sample size n increases without bound both the ECCE-MAD and the ECCE-R converge to 0 if the empirical cumulative distribution function of the scores converges uniformly to a continuous cumulative distribution function. For an *imperfectly* calibrated underlying distribution, Corollaries 9 and 10 show that the expected values of both the ECE^1 and the ECE^2 stay bounded away from 0 as the maximum bin width becomes arbitrarily small, and Theorem 11 shows that the expected values of both the ECCE-MAD and the ECCE-R stay bounded away from 0 for all sufficiently large n if the empirical cumulative distribution function of the scores converges uniformly to a cumulative distribution function. Thus, both the ECCE-MAD and the ECCE-R can distinguish every imperfectly calibrated underlying distribution from perfect calibration, given enough observed scores and corresponding responses, while in contrast there are imperfectly calibrated distributions which neither the ECE^1 nor the ECE^2 can distinguish from perfect calibration as the maximum width of a bin approaches 0, if the number of draws per bin remains bounded on some fixed range of scores.

Hence, any significance test based on the ECE^1 or the ECE^2 with a bounded number of draws per bin on some fixed range of scores has no power asymptotically for some alternatives or is asymptotically inconsistent. This exposes a fundamental trade-off inherent in the ECE^1 and the ECE^2: in order to attain nontrivial power and asymptotic consistency, the number of draws per bin must not stay bounded on any fixed range of scores, thus necessarily squandering observations that otherwise could have contributed additional power to the significance test (whereas the ECCE-MAD and the ECCE-R have no such explicit trade-off). The trade-off becomes especially hard to handle when the number of observations is limited; the ECCE-MAD and the ECCE-R work well without requiring any hard decisions, whereas the ECE^1 and the ECE^2 depend on the choice of bins, and that choice makes a big difference even asymptotically, in the limit that the number of observations tends to infinity (with no obvious best setting for finitely many observations).

To emphasize: to attain asymptotic consistency together with nontrivial asymptotic power against the fixed alternative distributions discussed above, both the ECE^1 and the ECE^2 require

infinitely many draws per bin for almost all bins (where "almost all" refers to "almost everywhere" on the unit interval $[0, 1]$)—requiring a density that is higher almost everywhere by an unbounded factor greater than what the ECCE requires.

2.5 Graphical methods

This subsection reviews the primary motivations for the definitions of the ECEs and the ECCEs—the ECEs and the ECCEs are scalar summary statistics for certain graphical methods of assessing calibration reviewed here.

2.5.1 Motivation for the empirical calibration error

Formulae (1) and (2) above express the average score and average response in bin j as

$$\tilde{S}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} S_j^k, \tag{11}$$

and

$$\tilde{R}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} R_j^k \tag{12}$$

for $j = 1, 2, \ldots, m$, respectively. Due to the central limit theorem, as n_j tends to infinity, $\tilde{R}_j$ concentrates around its expected value,

$$\mathbb{E}[\tilde{R}_j] = \frac{1}{n_j} \sum_{k=1}^{n_j} \mathbb{E}[R_j^k], \tag{13}$$

which is the average of the expected values of the responses in the bin. For a perfectly calibrated underlying distribution, $\mathbb{E}[R_j^k] = S_j^k$, and so (13) would be equal to (11). The difference from perfect calibration in bin j is therefore the difference between $\tilde{R}_j$ from (12) and $\tilde{S}_j$ from (11), in the limit of n_j tending to infinity. A so-called "calibration plot" or "reliability diagram" plots $\tilde{R}_j$ versus $\tilde{S}_j$ together with the (diagonal) line through a graph of $\tilde{S}_j$ versus $\tilde{S}_j$ for $j = 1, 2, \ldots, m$, so that the difference from perfect calibration is the vertical distance between the two graphs. Section 3 presents many examples of such reliability diagrams; see, for example, Figures 3, 4, 5, 6, 12, 13, 16, 17, 20, 21, 24, 25, 28, 29, 32, and 33.

The ECE^2 from (3) is the sum over $j = 1, 2, \ldots, m$ of the bin width $(S_{j+1}^1 - S_j^1)$ times the square of the difference between $\tilde{R}_j$ and $\tilde{S}_j$, where the latter difference approaches the expected amount of miscalibration in bin j in the limit of n_j tending to infinity; similarly, the ECE^1 from (4) is the sum over $j = 1, 2, \ldots, m$ of the bin width $(S_{j+1}^1 - S_j^1)$ times the absolute value of the difference between $\tilde{R}_j$ and $\tilde{S}_j$. In the limit that the bin width $(S_{j+1}^1 - S_j^1)$ tends to 0 uniformly over $j = 0, 1, \ldots, m$, and n_j tends to infinity uniformly, the ECE^1 becomes the total area between the graph of $\tilde{R}_j$ versus $\tilde{S}_j$ and the graph of $\tilde{S}_j$ versus itself, assuming that "area" is well-defined in terms of the Riemann sum (4), that is, that the ECE^1 converges to a unique limit. This is the case when there exists a fixed Riemann integrable function r such that $\mathbb{E}[R_j^k(n)] = r(S_j^k(n))$ for all $n = 1, 2, 3, \ldots$; $j = 1, 2, \ldots, m(n)$; $k = 1, 2, \ldots, n_j(n)$; again assuming that the bin width $(S_{j+1}^1 - S_j^1)$ tends to 0 uniformly over $j = 1, 2, \ldots, m$, and n_j tends to infinity uniformly over j.

2.5.2 Motivation for the empirical cumulative calibration error

Recall the definition in (5) of the cumulative differences:

$$C_k = \frac{1}{n} \sum_{j=1}^{k} (R_j - S_j) \tag{14}$$

for $k = 1, 2, \ldots, n$. Plotting the cumulative sum C_k versus k/n results in a graph whose expected slope is the amount of miscalibration; indeed, the expected slope from $j = k - 1$ to $j = k$ is

$$\frac{\mathbb{E}[C_k - C_{k-1}]}{k/n - (k-1)/n} = \mathbb{E}[R_k] - S_k \tag{15}$$

for $k = 1, 2, \ldots, n$—and $\mathbb{E}[R_k] - S_k$ is precisely the deviation from perfect calibration. Thus, the slope of a secant line connecting two points on the graph becomes the average miscalibration over the long range of k between the two points. Good calibration results in a flat, fairly horizontal graph that deviates little from zero. This motivates the definitions of the ECCE-MAD in (6) and of the ECCE-R in (7)—they measure the deviation from zero, that is, the deviation from a perfectly flat, horizontal graph of perfect calibration. The ECCE-MAD is simply the maximum absolute value of C_k, while the ECCE-R is simply the range of C_k, where the range is the maximum minus the minimum. As mentioned in Section 2.1.2 above, the absolute value of the total miscalibration $\sum_{j \in I} (R_j - S_j)/n$ over any interval I of indices is less than or equal to the ECCE-R—the ECCE-R is the maximum of the absolute value of the total miscalibration over any interval of indices: ECCE-R $= \max_I |\sum_{j \in I} (R_j - S_j)/n|$.

Note that slope is easy to perceive independently of any irrelevant constant vertical offset, and that slope in the plot of C_k versus k/n is equal to the amount of miscalibration (with the slope of a secant line becoming the average miscalibration over the scores between two distant points where the secant line intersects the graph). Section 3 presents many examples of such graphs of cumulative differences; see, for example, Figures 7, 14, 18, 22, 26, 30, and 34.

3. Results and discussion

This section illustrates the methods of the previous section via analysis of both synthetic and measured data sets.[3] The synthetic examples include the ground-truth known by construction. They first highlight practical problems with the ECEs, then validate the theory of the previous section directly and explicitly. The examples on measured data display even more extreme practical problems with the ECEs, especially in comparison with the ECCEs. Section 3.1 presents the synthetic examples, while Section 3.2 analyzes in detail one of the most popular data sets from computer vision, ImageNet of Russakovsky et al. (2015).

The following are minor details common to both Section 3.1 and Section 3.2:

1. "P-values" (also known as "attained significance levels") which are exact in the asymptotic limit that the sample size n tends to infinity accompany every value for the ECCE-MAD and for the ECCE-R reported in the captions of the figures; efficient methods for computing such P-values are detailed by Tygert (2022).

2. All reliability diagrams displayed in the present paper include "error bars" (actually lines, not bars) plotted in light gray. Each diagram includes 20 such light-gray graphs, obtained via

3. Permissively licensed open-source software that automatically reproduces all figures and statistics reported below is available at `https://github.com/facebookresearch/ecevecce`

bootstrap resampling, corresponding to a confidence level of around 95%. Details on their computation are available in the appendix of Tygert (2021a).

3. When the bins are equispaced with respect to the scores (the scores are the predicted probabilities), we replace the bin width $(S^1_{j+1} - S^1_j)$ in (3) and (4) with $1/m$, where m is the number of bins; both (3) and (4) are still Riemann sums with this change, and so all the analysis given above remains valid without modification. Such replacement is canonical in the classical reliability diagrams and ECEs when the bins are equispaced.

3.1 Synthetic examples

This subsection presents the results of numerical experiments on a toy example, generated synthetically so that the complete ground-truth is known exactly. Knowing the ground-truth facilitates a fully rigorous evaluation and validation of the methods of the previous section. Figures 8–10 illustrate the theorems and corollaries of Section 2 as explicitly as possible, as detailed in the penultimate paragraph of the present subsection. Figures 1–7 set the stage, introducing the synthetic examples and some problems with binning encountered in practice.

Figure 1 displays the four kinds of empirical calibration errors, as a function of m, the number of bins. Whether any number m of bins is optimal (or even merely representative) is entirely unclear. The values of the empirical calibration errors vary widely as the number m of bins varies. Figure 1 corresponds to the sample size $n = 32{,}768$ used in the present subsection; Appendix B contains analogous plots for the samples sizes $n = 8{,}192$ and $n = 131{,}072$.

Figure 2 plots the probabilities of success for the Bernoulli distributions from which the synthetic data set draws responses at the specified scores, where the scores are equispaced in the upper plot of the figure, then squared in the middle plot, and finally square rooted from the original equispaced values in the lower plot. The sample size is $n = 32{,}768$, which is the number of scores (each paired with a response drawn from the Bernoulli distribution whose probability of success is graphed) for each plot.

Figures 3–6 display the reliability diagrams from Section 2.5.1 (both with bins that are roughly equispaced with respect to the scores and with each bin containing the same number of observations), for $m = 16$ bins and $m = 64$. Figure 7 displays the cumulative plot from Section 2.5.2, along with the ground-truth ideal, constructing the ideal graph using the exact expected values of the Bernoulli distributions from which the observed responses are drawn; the empirical plot closely resembles the ground-truth ideal.

Figures 8–10 illustrate explicitly the theory of Section 2. The upper plots of Figure 8 correspond to Corollaries 2 and 4, while the upper plots of Figure 9 correspond to Corollary 7; the lower plots of Figure 8 correspond to Corollaries 9 and 10, while the lower plots of Figure 9 correspond to Theorem 11. Figure 10 displays the ECCE-MAD and the ECCE-R both normalized by σ_n from (9); the perfectly calibrated data of the upper plots in Figure 10 results in the ECCE-MAD / σ_n hovering around its asymptotic expected value $\sqrt{\pi/2} \approx 1.2533$ (asymptotic in the limit of the sample size n tending to infinity) and in the ECCE-R / σ_n hovering around its asymptotic expected value $2\sqrt{2/\pi} \approx 1.5958$. Derivations of these expected values in the limit of the sample size n tending to infinity are available in Section 3 of Tygert (2022). Figure 8 displays graphically how the ECEs hit a noise floor, staying around the same value for both the perfectly and imperfectly calibrated distributions, irrespective of the number of observations. In contrast, Figure 9 illustrates how the ECCEs approach 0 rapidly for the perfectly calibrated distribution as the sample size n increases, while staying well away from 0 for the imperfectly calibrated distribution. Thus, the ECEs have trouble telling apart the perfectly and imperfectly calibrated distributions, whereas the power of the ECCEs increases indefinitely as the sample size grows.

The captions of the figures discuss the results and their consequences.

3.2 ImageNet

This subsection presents the results of numerical experiments on the standard training data set "ImageNet-1000" of Russakovsky et al. (2015), which is a very popular data set in computer vision.

The standard training data set "ImageNet-1000" contains a thousand labeled classes, each containing about 1,300 images corresponding to a particular noun (often an animal such as a "night snake," a "sidewinder or horned rattlesnake," or an "Eskimo dog or husky"); the total number of images in the data set is n = 1,281,167. To generate the corresponding plots, we calculate the scores $S_1, S_2, \ldots, S_n$ using the pretrained ResNet18 classifier of He et al. (2016) from the computer-vision module, "torchvision," in the PyTorch software library of Paszke et al. (2019); the score for an image is the probability assigned by the classifier to the class predicted to be most likely, with the scores randomly perturbed by about one part in 10^8 to guarantee their uniqueness. For $k = 1, 2, \ldots, n$, the response R_k corresponding to a score S_k is $R_k = 1$ when the class predicted to be most likely is the correct class, and $R_k = 0$ otherwise.

The figures presented below consider both the subsets of the full data set corresponding to individual classes as well as the full data set with all classes simultaneously. Figures 11–30 pertain to individual classes, while Figures 31–34 pertain to all classes together.

Figures 11–14 consider the class corresponding to the night snake, Figures 15–18 consider the sidewinder or horned rattlesnake, Figures 19–22 consider the Eskimo dog or husky, Figures 23–26 consider the wild boar, and Figures 27–30 consider sunglasses. In each of these sets of four figures, the first figure displays the four kinds of empirical calibration errors—the ECE^1 and the ECE^2 for when the bins are equispaced along the scores, and the ECE^1 and the ECE^2 for when each bin (aside from the last) contains the same number of observations. The second figure in the set of four provides examples of the reliability diagrams from Section 2.5.1, with the bins chosen such that $S_1^1, S_2^1, \ldots, S_m^1$ are roughly equispaced on the unit interval $[0, 1]$ (with $m = 8$ in the upper plot and $m = 32$ in the lower plot). The third figure in each set of four gives examples of the reliability diagrams from Section 2.5.1, now with the bins chosen such that $n_1 = n_2 = \cdots = n_{m-1} \approx n_m$ (again with $m = 8$ in the upper plot and $m = 32$ in the lower plot). The fourth figure in the set of four provides an example of the cumulative plot from Section 2.5.2. For each of Figures 11–30, the sample size is n = 1,300.

Figure 31 displays the empirical calibration errors for all classes analyzed simultaneously, so that the sample size n = 1,281,167. Figure 32 gives examples of the reliability diagrams from Section 2.5.1 for all 1,000 classes together, with the bins chosen such that $S_1^1, S_2^1, \ldots, S_m^1$ are roughly equispaced on the unit interval $[0, 1]$ (with $m = 128$ in the upper plot and m = 1,024 in the lower plot). Figure 33 provides examples of the reliability diagrams from Section 2.5.1 for all 1,000 classes simultaneously, with the bins chosen such that $n_1 = n_2 = \cdots = n_{m-1} \approx n_m$ (again with $m = 128$ in the upper plot and m = 1,024 in the lower plot). Figure 34 provides an example of the cumulative plot from Section 2.5.2 for all classes together. Only the cumulative plot (Figure 34) conveniently reveals that a third of all observations (specifically, those with probabilities of at least 0.97) are well-calibrated. The ranges of the graphs in Figure 31 are relatively narrow as m, the number of bins, varies through the values 8, 16, 32, ..., 1,024; the empirical calibration errors could plausibly constitute decent metrics in this setting, on account of their relatively stable values as m varies through the values 8, 16, 32, ..., 1,024.

In contrast, all the empirical calibration errors vary enormously as a function of m, the number of bins, for every individual class from ImageNet investigated here—the range of the graphs in every one of Figures 11, 15, 19, 23, and 27 is wide even with merely modest variations in m. Which

choice of m, the number of bins, and binning strategy is best—if any—is entirely unclear from these plots. Whether any choice of m or binning strategy yields a good metric must be seriously questionable, as the choice makes such a big difference in the value of the metric. Which choices are best is unclear.

Thus, the empirical calibration errors may be most meaningful when the probabilities of success for the Bernoulli distributions underlying the observed data are smooth as a function of m, the number of bins, and the sample size n greatly exceeds the minimum required to assess statistical significance reliably by averaging away noise from sampling. If the probabilities of success display multiscale behavior as a function of m, with interesting variations present at finer and finer scales, then any choice of m will necessarily miss interesting variations or fail to perform enough averaging to distinguish signal from noise. In accord with the theory of Section 2, obtaining meaningful empirical calibration errors apparently requires the sample size n to be much larger than the ideal attained by the empirical *cumulative* calibration errors.

More detailed discussion is available in the captions of the figures.

4. Conclusion

A trade-off between statistical confidence and power to resolve variations as a function of score is inherent to the empirical calibration errors (ECEs) based on binning, while the empirical *cumulative* calibration errors (ECCEs) have no such explicit trade-off. The theory of Section 2 proves that this trade-off results in the ECEs requiring observations whose density is infinitely higher than what the ECCEs require to attain the same consistency and power against any fixed alternative distribution, where "infinitely higher" means asymptotically, in the limit of the sample size tending to infinity (or in the limit of the statistical confidence level tending to 100%). Consonant with the asymptotic theory, the examples of Section 3 illustrate at the finite sample sizes of greatest interest in practice the drastically higher density required by the ECEs, together with the ECEs' trade-off between confidence and resolving power. The ECEs also exhibit an extreme dependence on the choice of bins, with different choices of bins yielding significantly different values for the ECE metrics; choosing among the possible binnings can be confusing, yet makes all the difference. In contrast, the ECCEs yield trustworthy results without needing such large numbers of observations and without needing to set any parameters. Furthermore, P-values (also known as "attained statistical significance levels") that are asymptotically perfectly-calibrated in the limit of the sample size tending to infinity are available for the ECCEs via the simple, convenient, efficient methods of Tygert (2022). All in all, the ECEs are unreliable and largely unusable, while the ECCEs are reliable and easy to use.

Acknowledgments

We would like to thank Kamalika Chaudhuri and Isabel Kloumann.

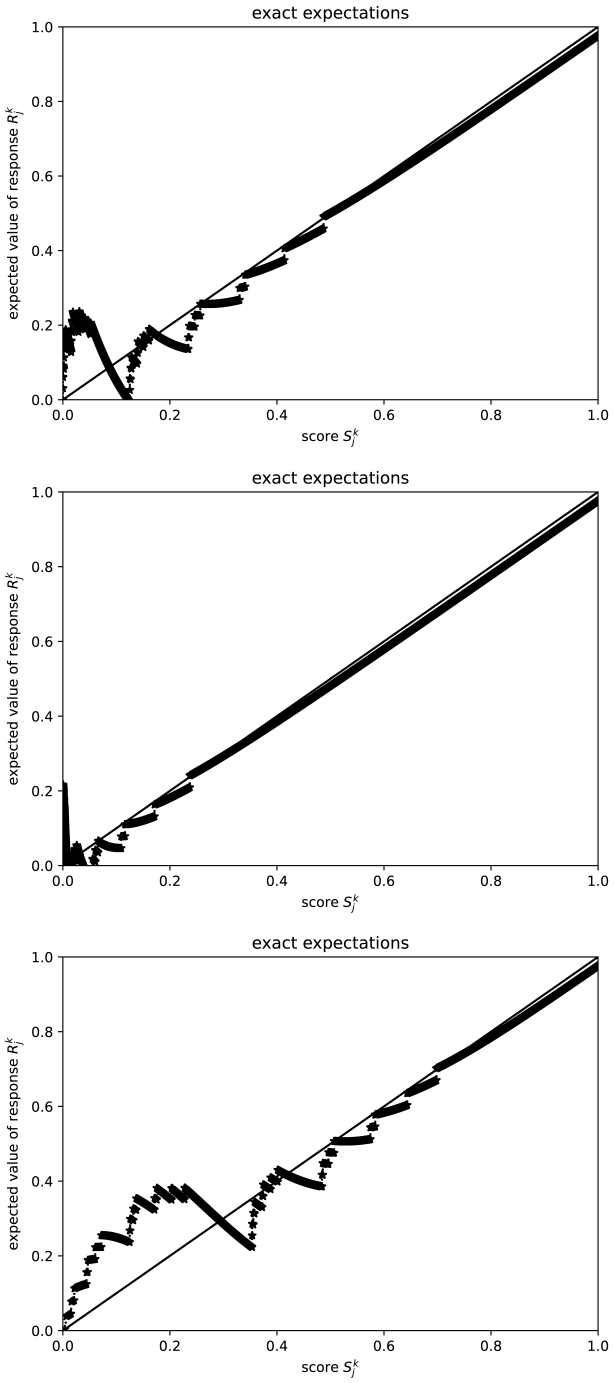

Figure 2: Probabilities of success for the Bernoulli distributions underlying the synthetic data set (which takes independent draws from these distributions to obtain the observed responses). The scores are equispaced in the top plot, squared in the middle plot, and square rooted in the bottom plot, with sample size $n = 32{,}768$.

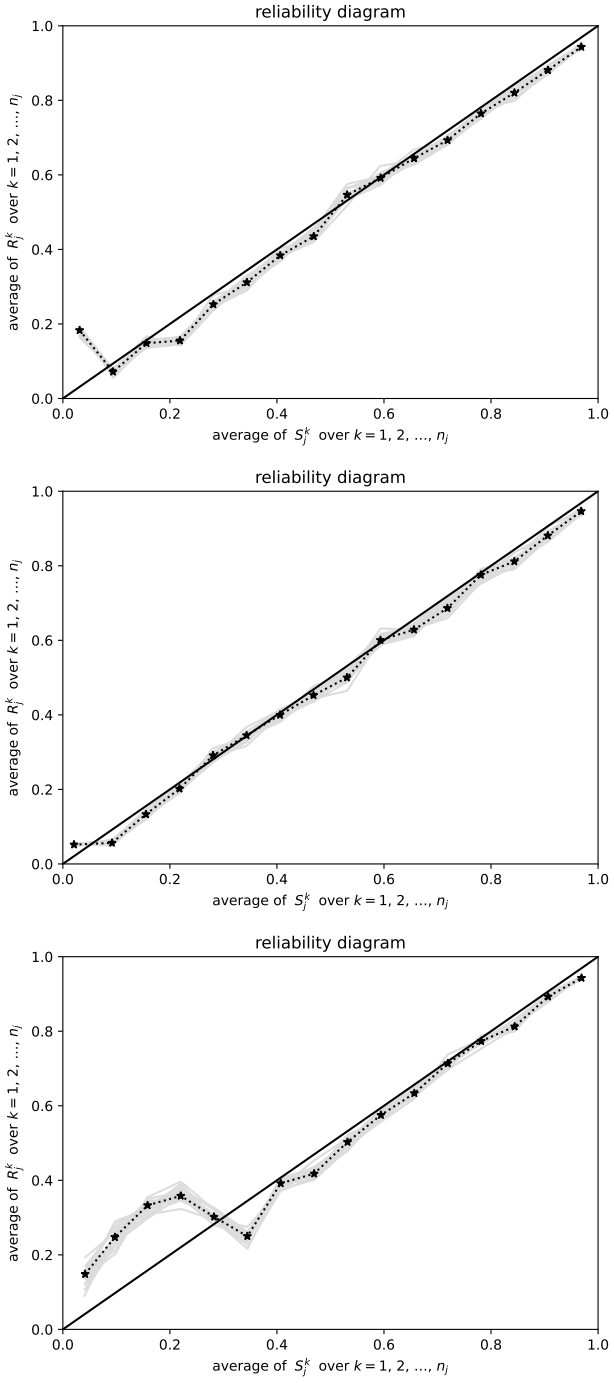

Figure 3: Reliability diagrams for the synthetic data set, with the bins roughly equispaced. The scores are equispaced in the top plot, squared in the middle plot, and square rooted in the bottom plot, with $m = 16$ bins and sample size $n = 32{,}768$.

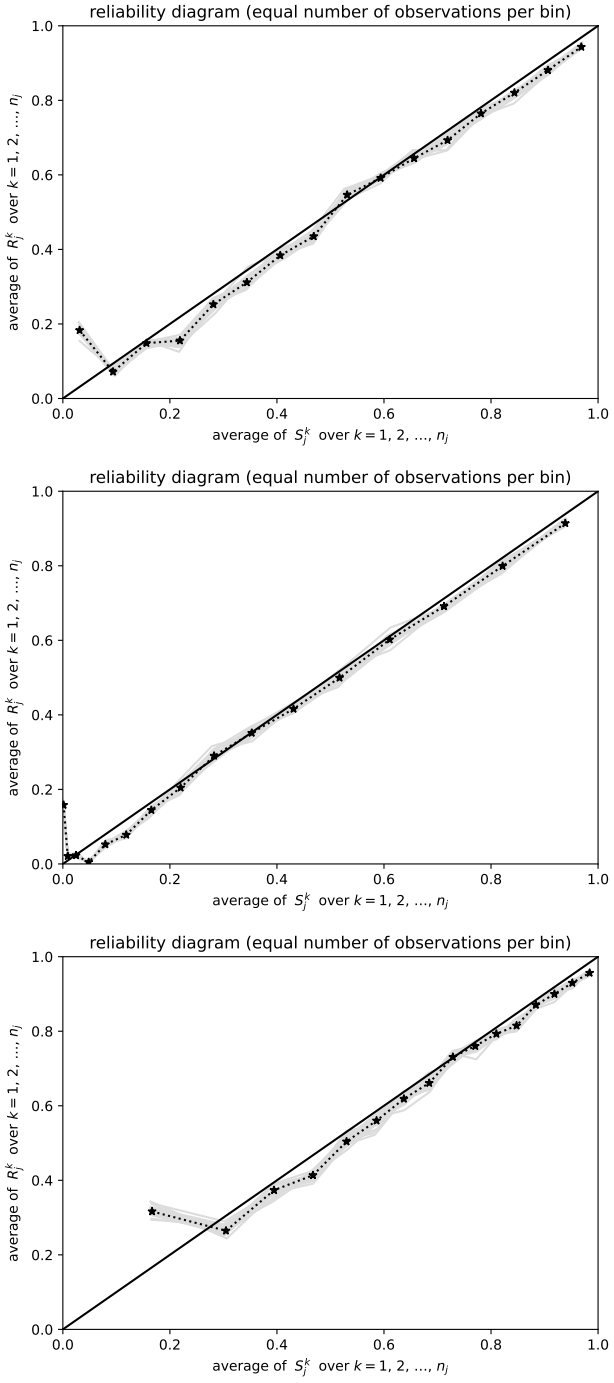

Figure 4: Reliability diagrams for the synthetic data set, with an equal number of observations per bin. The scores are equispaced in the top plot, squared in the middle plot, and square rooted in the bottom plot, with $m = 16$ bins and sample size $n = 32{,}768$.

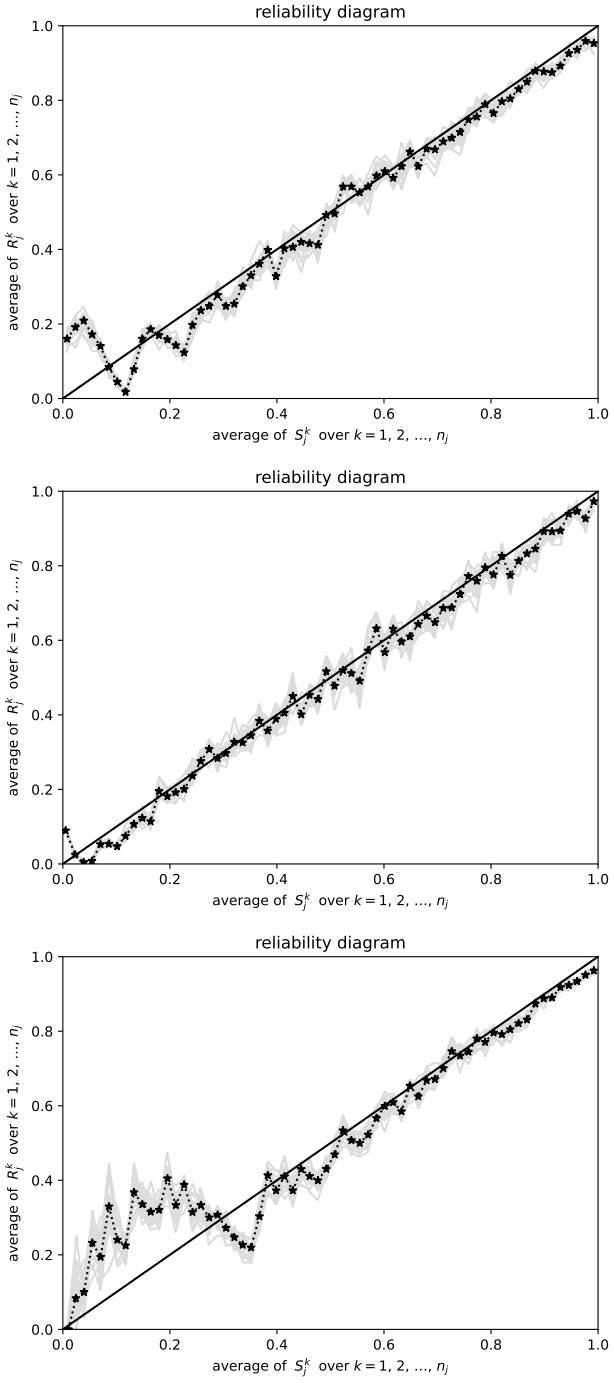

Figure 5: Reliability diagrams for the synthetic data set, with the bins roughly equispaced. The scores are equispaced in the top plot, squared in the middle plot, and square rooted in the bottom plot, with $m = 64$ bins and sample size $n = 32{,}768$.

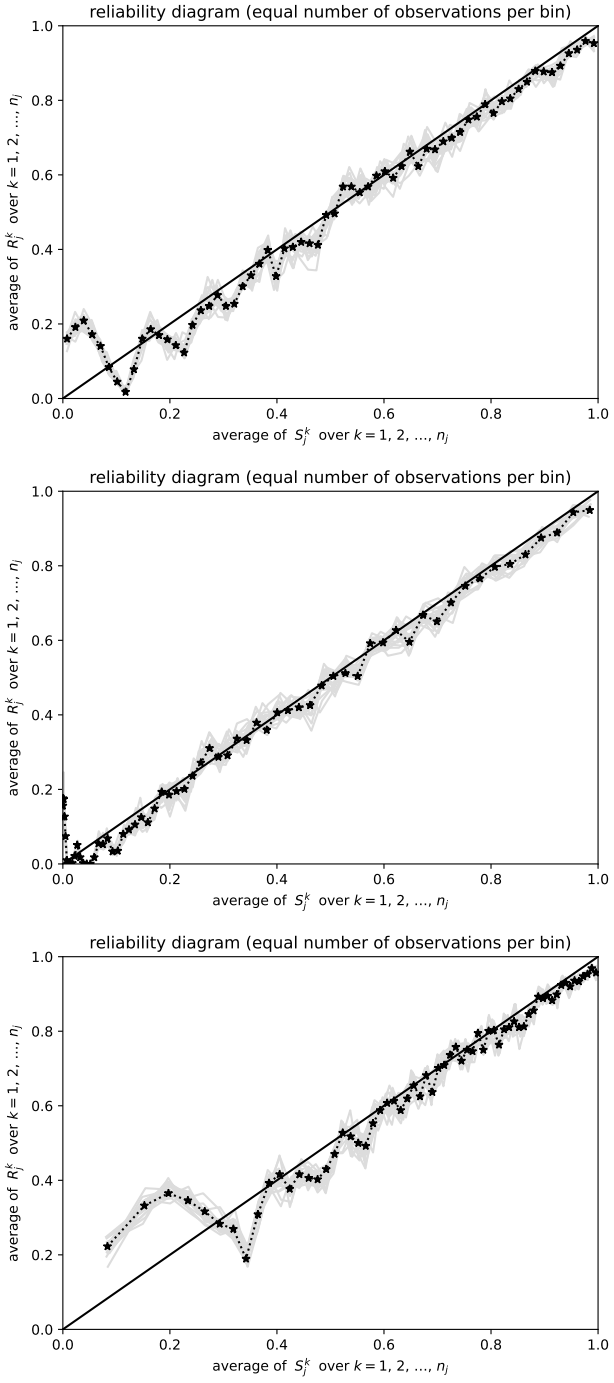

Figure 6: Reliability diagrams for the synthetic data set, with an equal number of observations per bin. The scores are equispaced in the top plot, squared in the middle plot, and square rooted in the bottom plot, with $m = 64$ bins and sample size $n = 32{,}768$

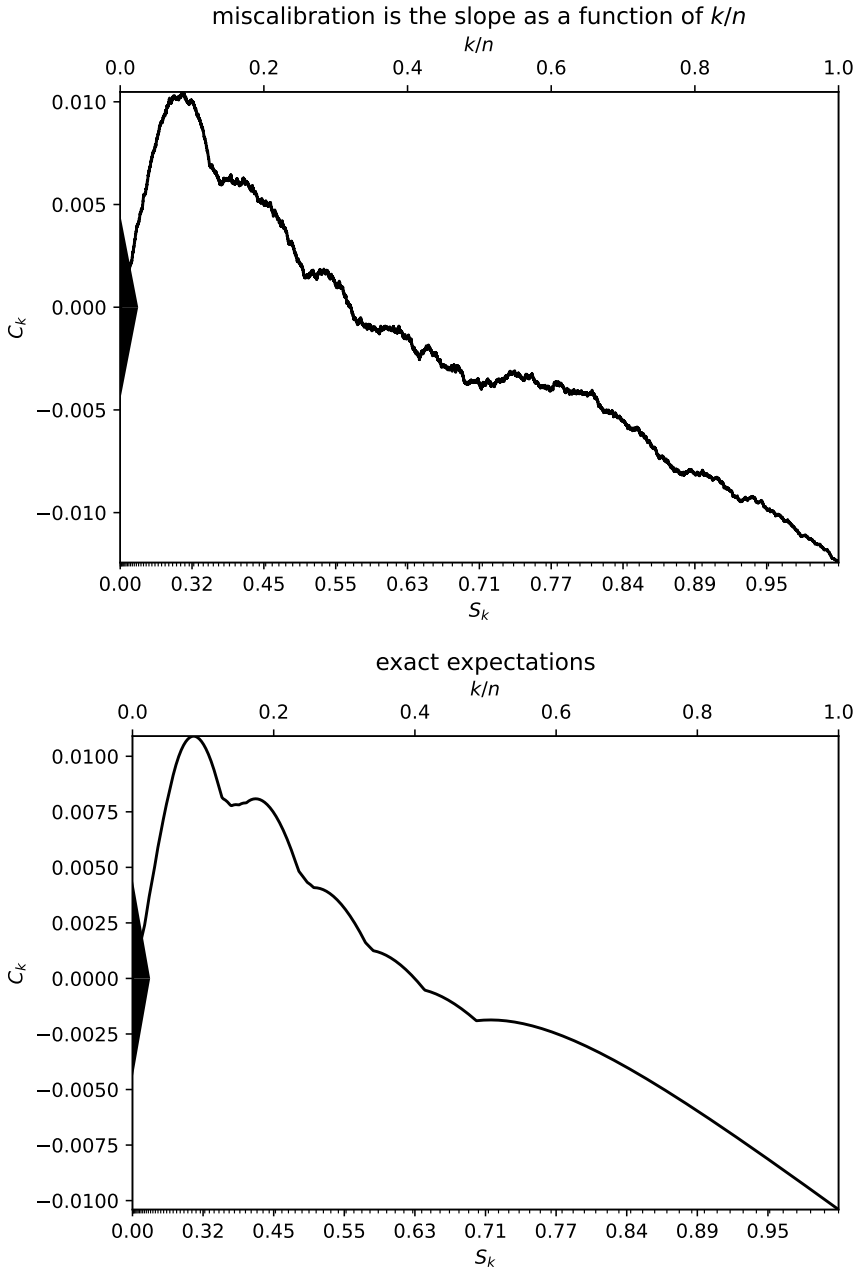

Figure 7: Cumulative plot for the synthetic data set with square-rooted scores and sample size $n =$ 32,768. The ECCE-MAD is $0.01243/\sigma_n = 5.512$, and the ECCE-R is $0.02291/\sigma_n = 10.16$; the associated asymptotic P-values are 7.1E–08 and zero to double-precision accuracy, respectively. The upper plot is based on the empirical observations, while the lower plot is the ideal, based on full knowledge of the exact probabilities of success for the Bernoulli distributions from which the empirical observations were drawn. The slopes of secant lines in the upper plot appear to match the slopes of the corresponding secants in the lower plot reasonably well, aside from the expected statistical fluctuations (whose expected standard deviation is a quarter of the height of the triangle at the origin).

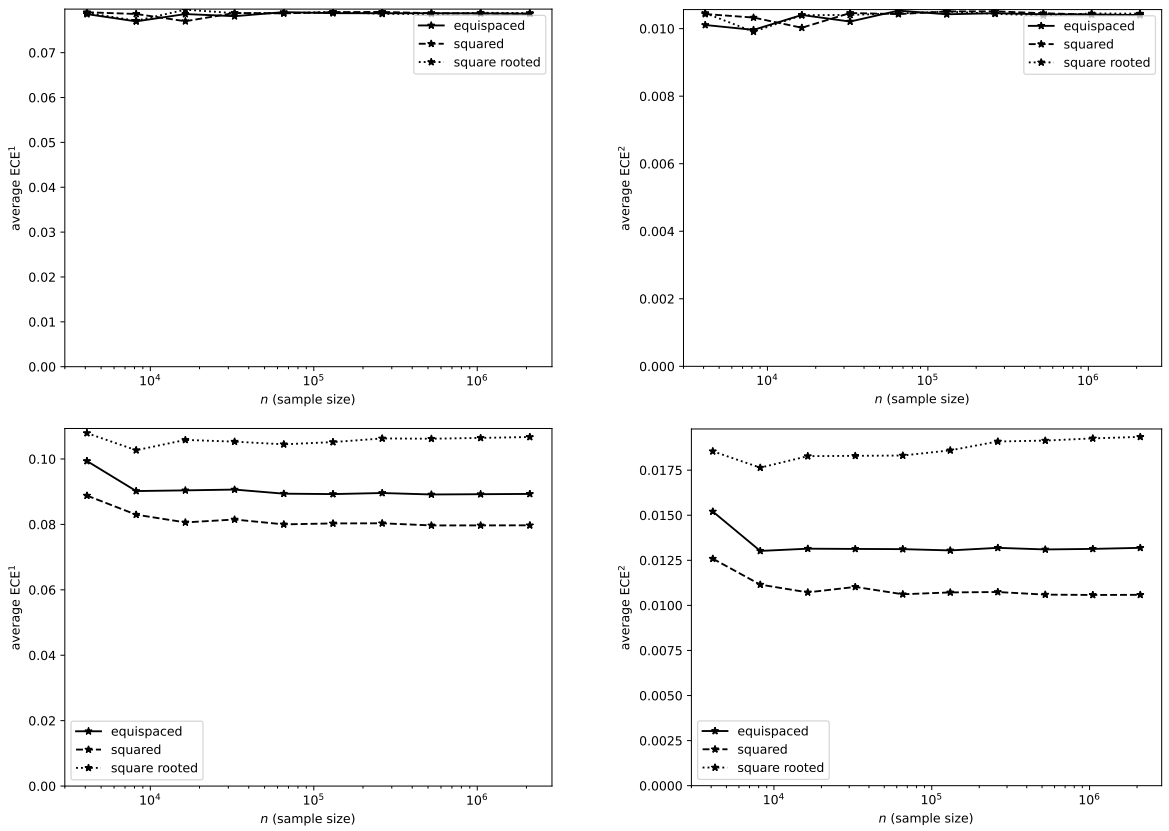

Figure 8: The upper plots display the ECE^1 and the ECE^2 averaged over 9 synthetic data sets (reducing random variations by about $\sqrt{9} = 3$), each of which is perfectly calibrated. The lower plots display the ECE^1 and the ECE^2 averaged over 9 synthetic data sets, each drawn from the distribution depicted in Figure 2 for the sample size $n = 32{,}768$. In all cases, each bin contains the same number of observations, namely 16; so m, the number of bins, is $n/16$. The scores are equispaced, equispaced then squared, or equispaced and then square rooted, as indicated in the legends for the plots; the underlying alternative distributions of responses used for the lower plots here which correspond to these different distributions of scores are the upper, middle, and lower plots of Figure 2 for $n = 32{,}768$, respectively. Notice how all values for the ECE^1 are quite similar, as are all values for the ECE^2; distinguishing the perfectly calibrated data sets from the alternative distribution of Figure 2 is very hard.

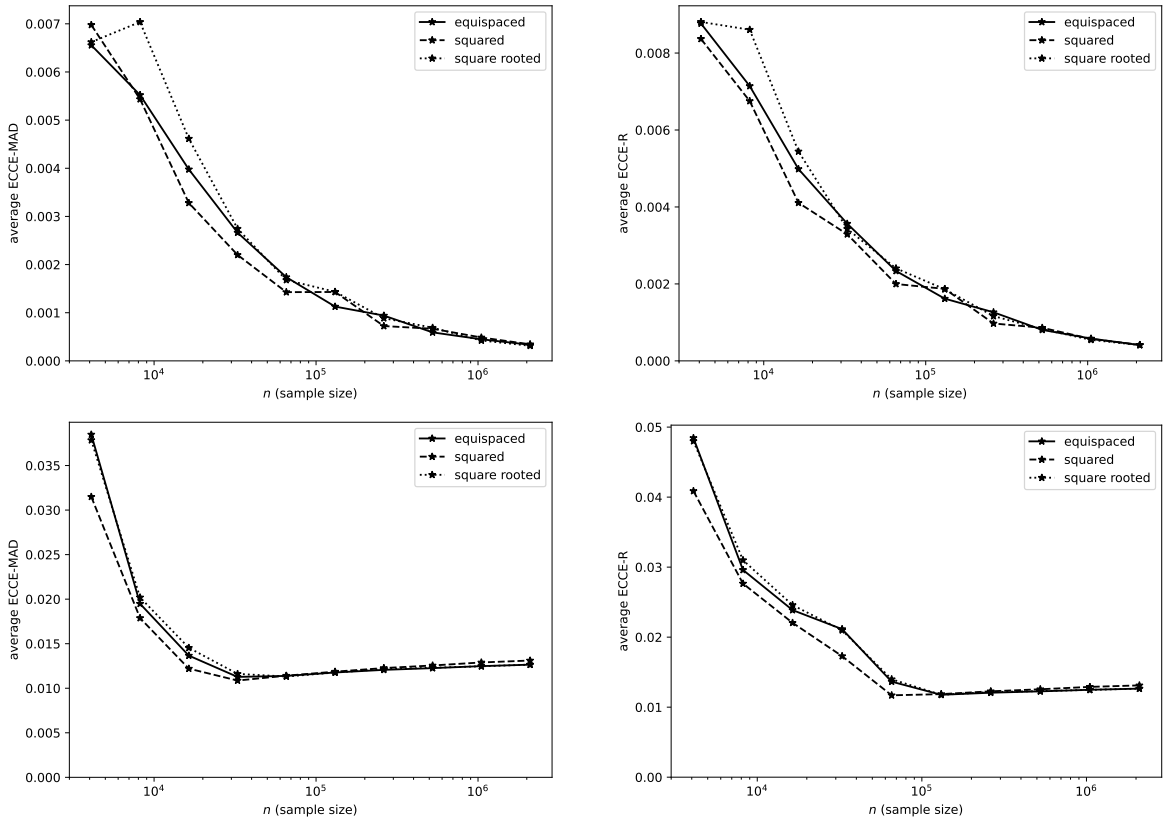

Figure 9: The upper plots display the ECCE-MAD and the ECCE-R averaged over 9 synthetic data sets (hence reducing random variations by a factor of about $\sqrt{9} = 3$), each of which is perfectly calibrated. The lower plots display the ECCE-MAD and the ECCE-R averaged over 9 synthetic data sets, each drawn from the distribution depicted in Figure 2 for the sample size $n = 32{,}768$. The scores are equispaced, equispaced then squared, or equispaced and then square rooted, as indicated in the legends for the plots; the underlying alternative distributions of responses used in the lower plots here which correspond to these different distributions of scores are the upper, middle, and lower plots of Figure 2 for $n = 32{,}768$, respectively. Notice how the values for the ECCE-MAD get much, much lower in the upper plot than in the lower plot, and, similarly, how the values for the ECCE-R get much, much lower in the upper plot than in the lower plot; distinguishing the perfectly calibrated data sets from the alternative distribution of Figure 2 is easy with the ECCE-MAD or the ECCE-R, with high statistical confidence that increases as the sample size n becomes large. The graphs in the lower plots stay flat as n becomes large, while the graphs in the upper plots decay rapidly.

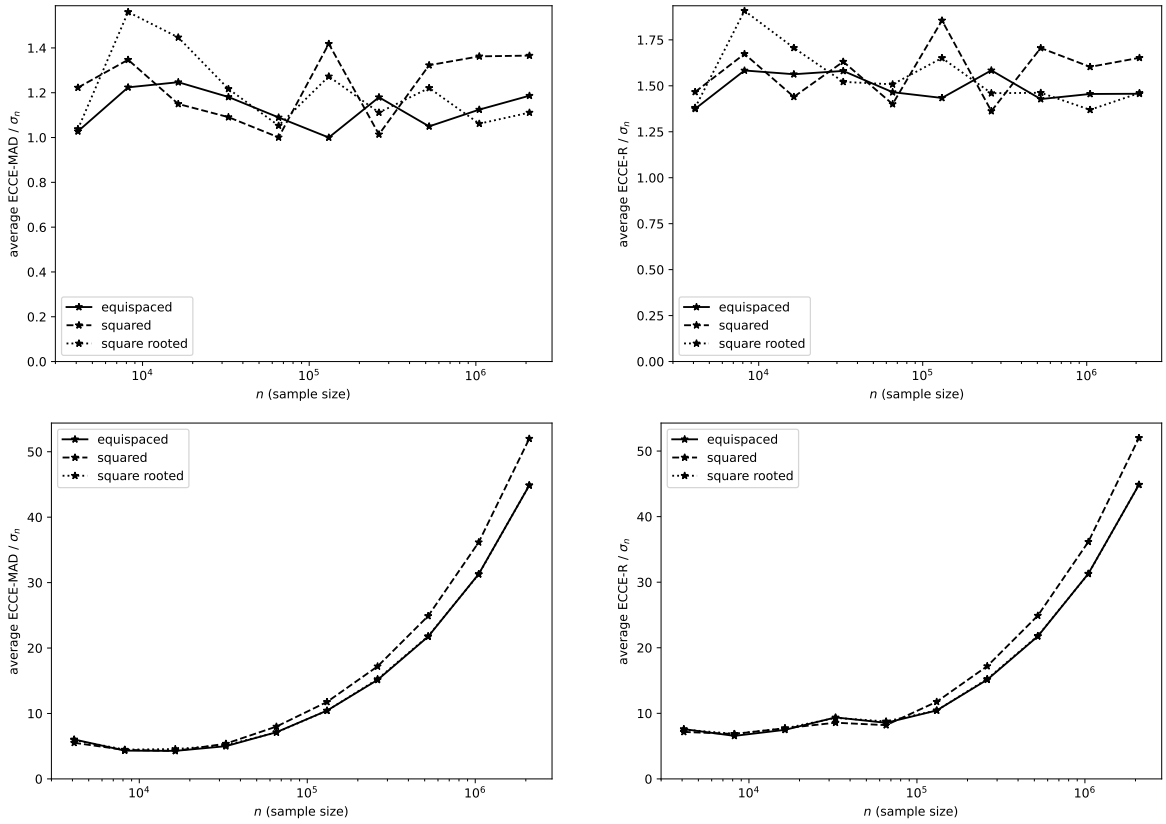

Figure 10: The upper plots display the normalized ECCE-MAD and the normalized ECCE-R averaged over 9 synthetic data sets (which reduces random variations by a factor of around $\sqrt{9} = 3$), each of which is perfectly calibrated. The lower plots display the normalized ECCE-MAD and the normalized ECCE-R averaged over 9 synthetic data sets, each drawn from the distribution depicted in Figure 2 for the sample size $n = 32{,}768$. The scores are equispaced, equispaced then squared, or equispaced and then square rooted, as indicated in the legends for the plots; the underlying alternative distributions of responses used for the lower plots here which correspond to these different distributions of scores are the upper, middle, and lower plots of Figure 2 for $n = 32{,}768$, respectively. The normalization factor σ_n is defined in (9). The average across the 9 realizations is over the quotient of the ECCE by σ_n, with a different value of σ_n for every realization. Notice how the values for the ECCE-MAD / σ_n get much, much higher in the lower plot than in the upper plot, and, similarly, how the values for the ECCE-R / σ_n get much, much higher in the lower plot than in the upper plot; distinguishing the perfectly calibrated data sets from the alternative distribution of Figure 2 is easy with the normalized ECCE-MAD or the normalized ECCE-R, with high statistical confidence that increases as the sample size n becomes large. The graphs in the upper plots stay flat as n becomes large, while the graphs in the lower plots increase explosively.

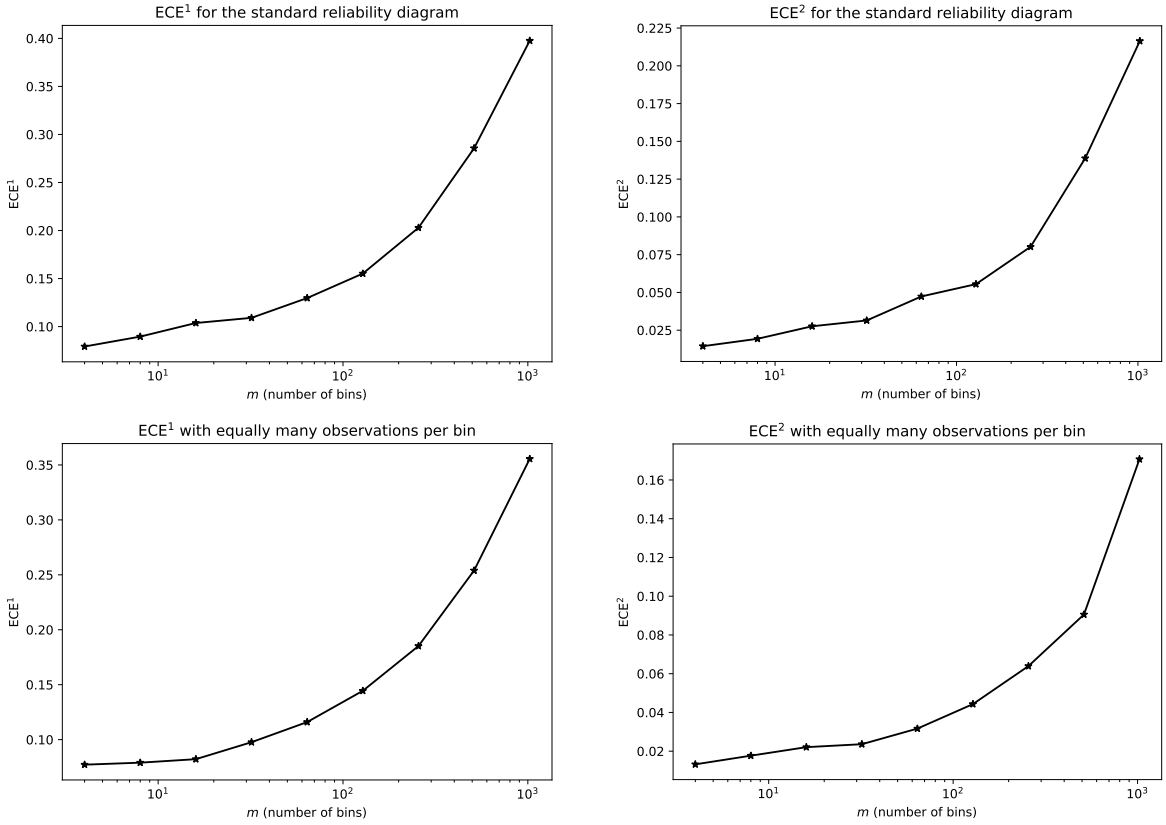

Figure 11: Empirical calibration errors for the night snake (*Hypsiglena torquata*), with sample size $n = 1{,}300$.

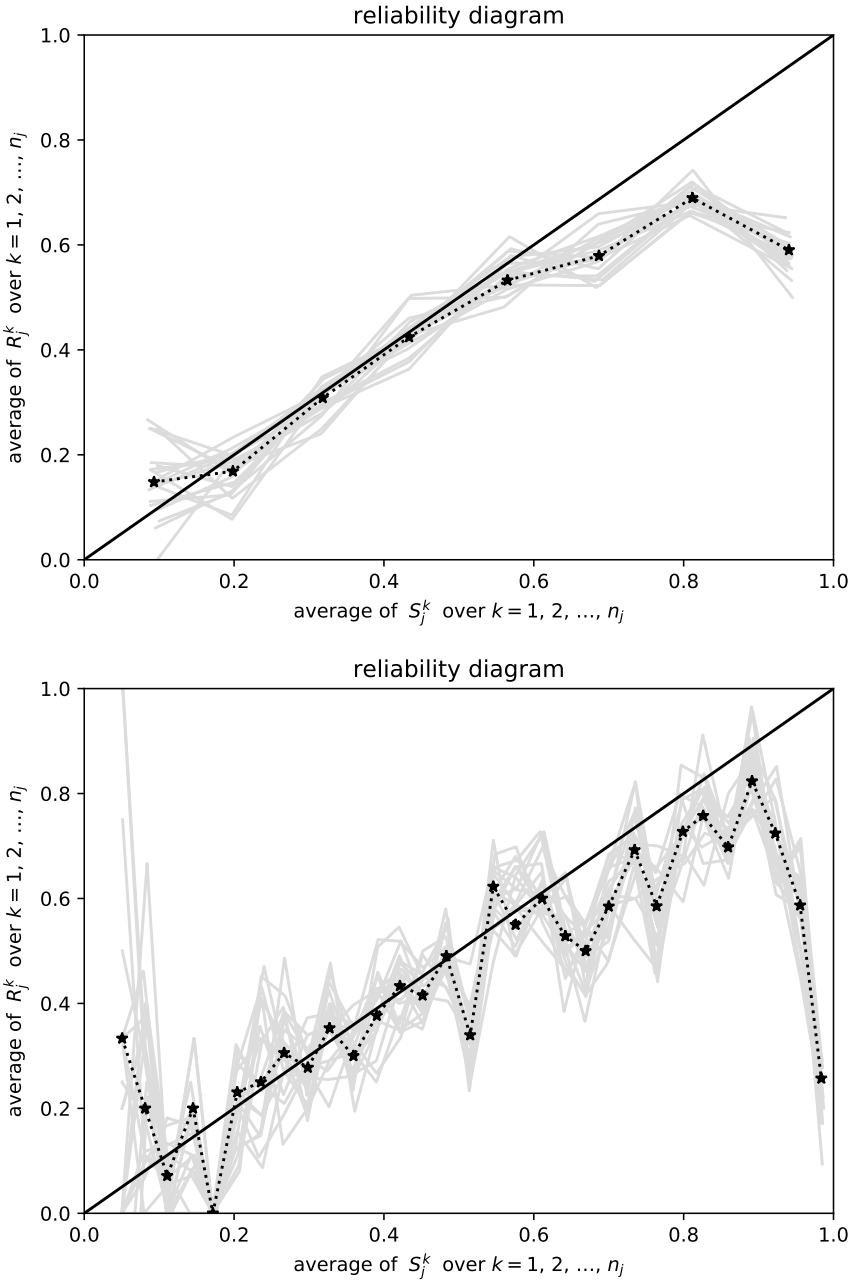

Figure 12: Reliability diagrams for the night snake (*Hypsiglena torquata*), with the bins roughly equispaced. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

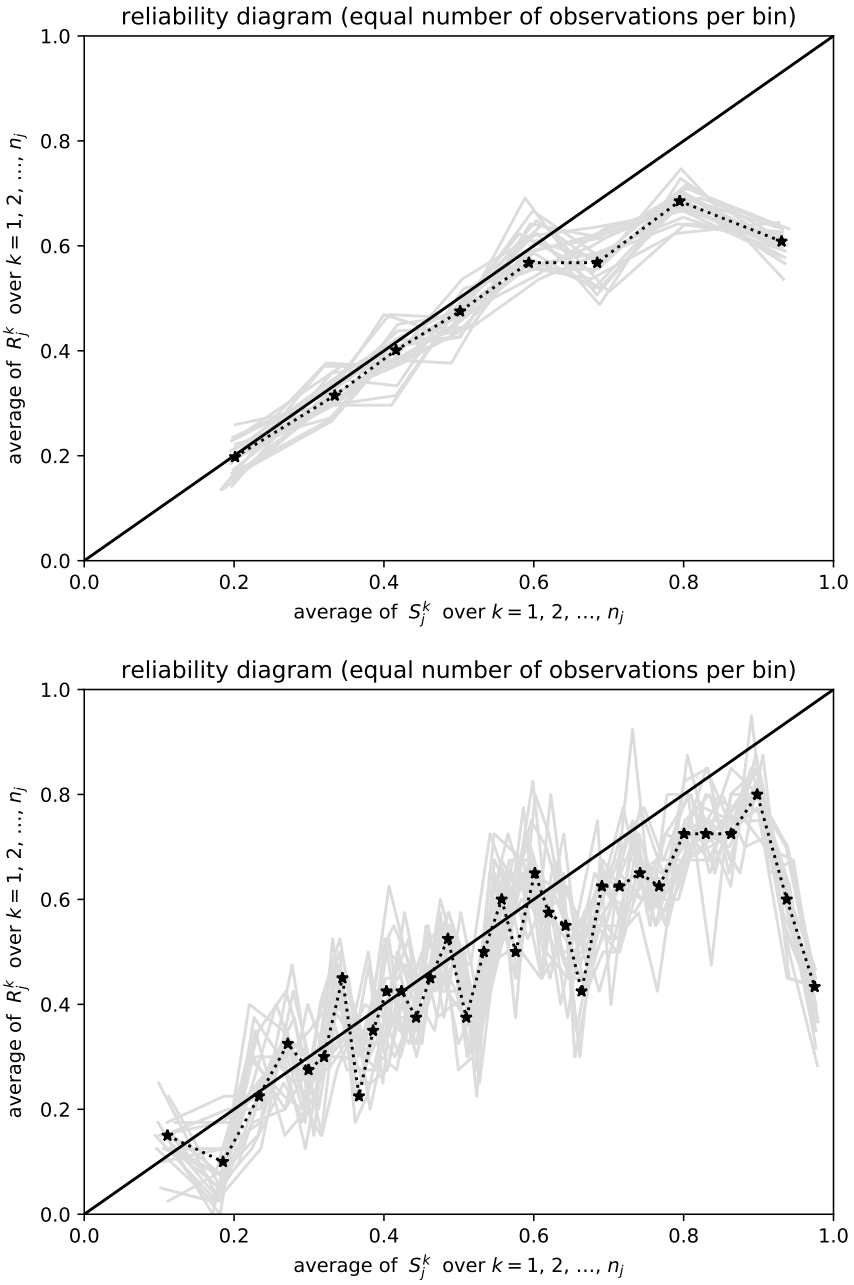

Figure 13: Reliability diagrams for the night snake (*Hypsiglena torquata*), with an equal number of observations per bin. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

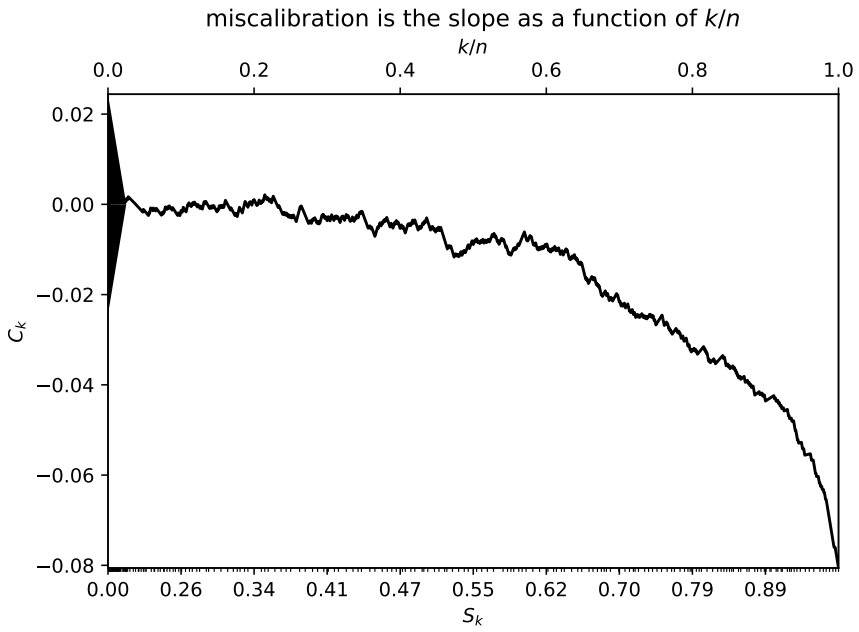

Figure 14: Cumulative plot for the night snake (*Hypsiglena torquata*), with sample size $n = 1{,}300$. The ECCE-MAD is $0.08059/\sigma_n = 6.607$, and the ECCE-R is $0.08270/\sigma_n = 6.780$; the associated asymptotic P-values are 7.8E–11 and 4.8E–11, respectively.

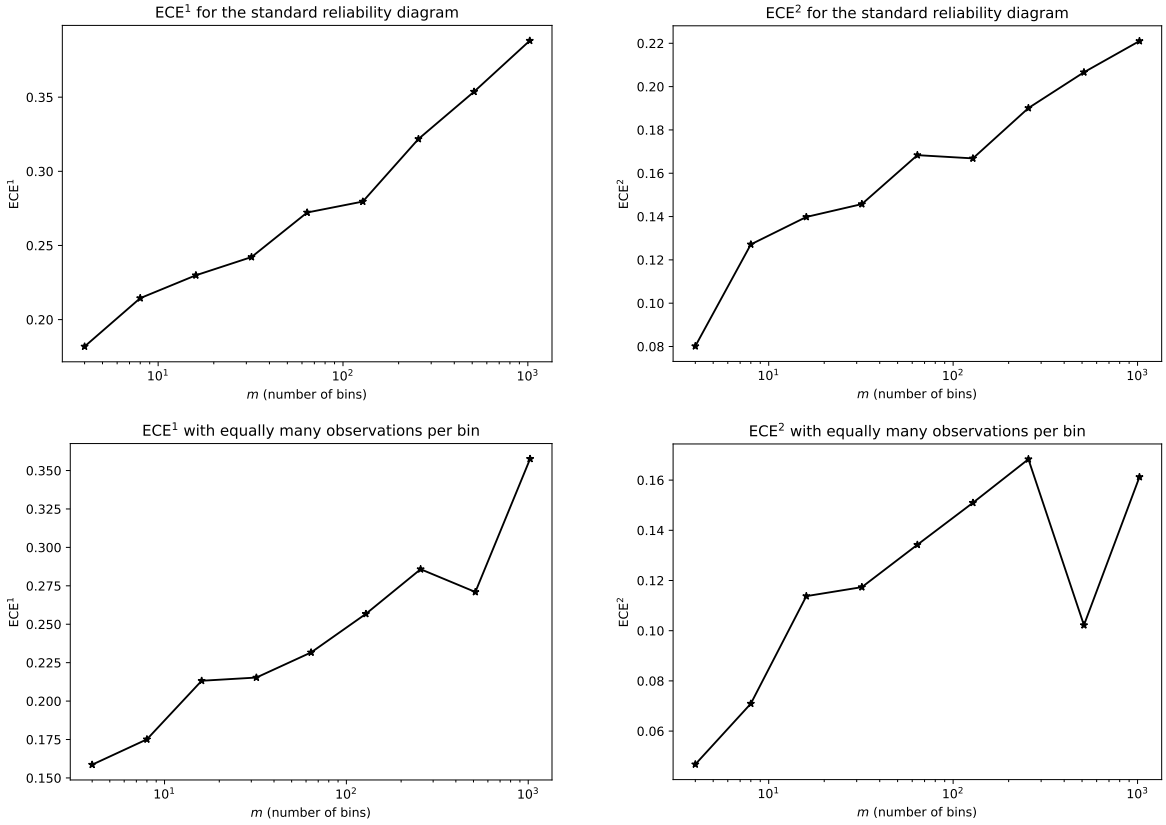

Figure 15: Empirical calibration errors for the sidewinder or horned rattlesnake (*Crotalus cerastes*), with sample size $n = 1{,}300$.

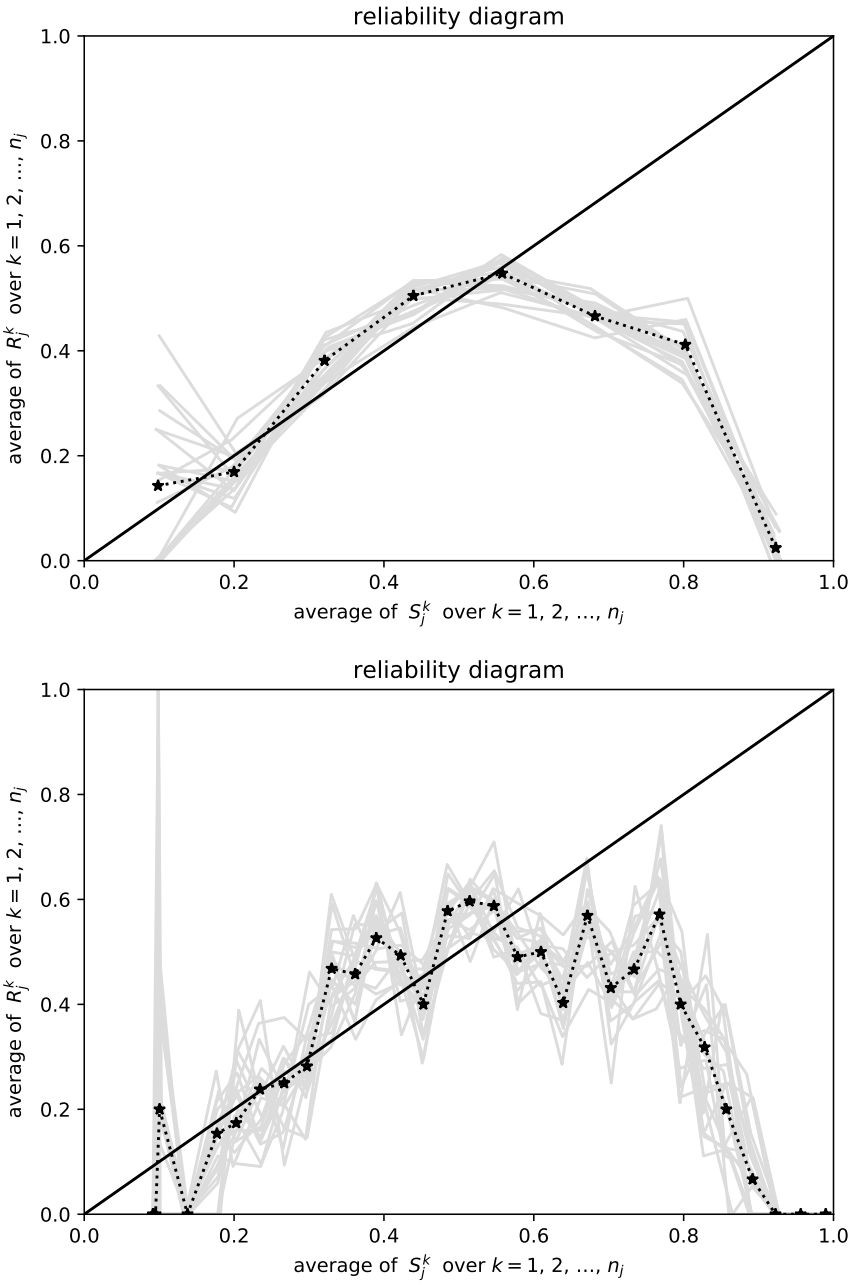

Figure 16: Reliability diagrams for the sidewinder or horned rattlesnake (*Crotalus cerastes*), with the bins roughly equispaced. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

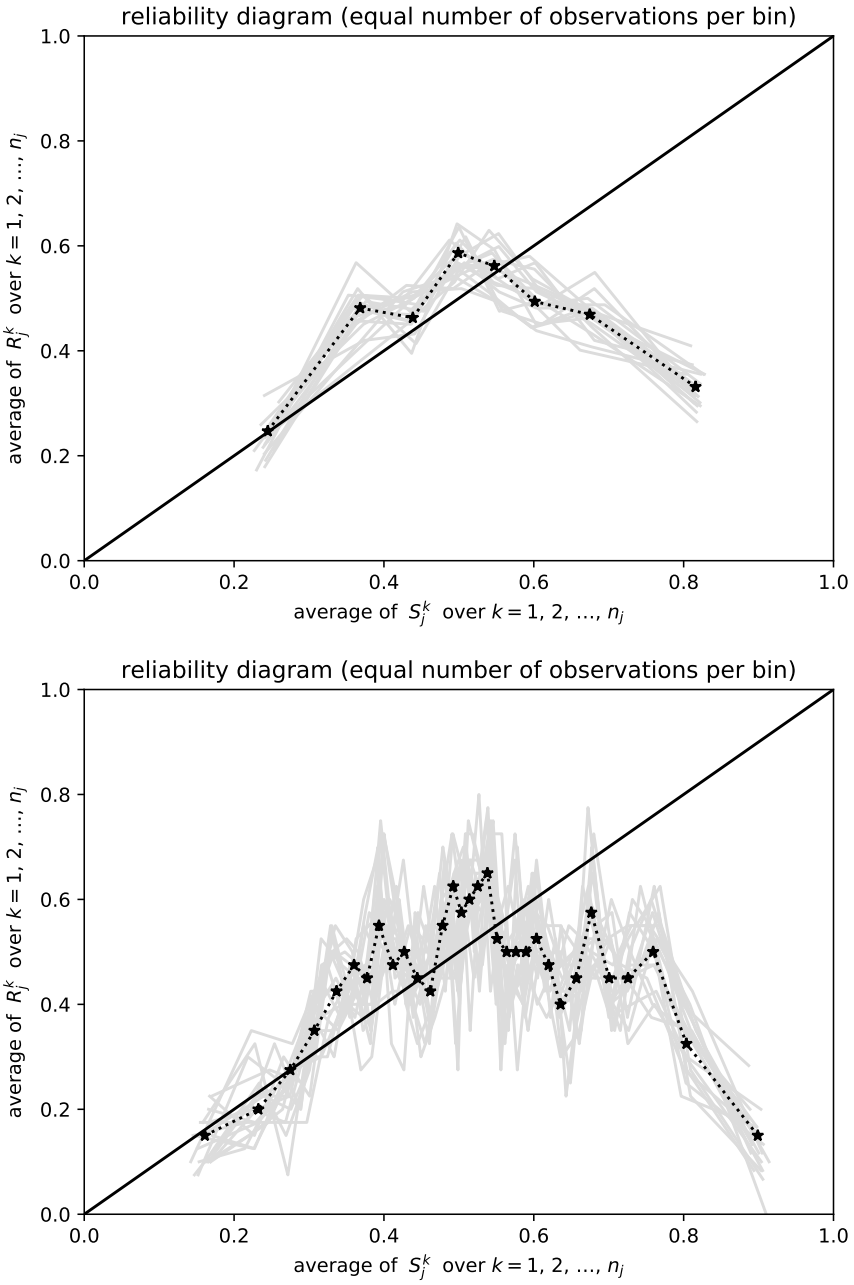

Figure 17: Reliability diagrams for the sidewinder or horned rattlesnake (*Crotalus cerastes*), with an equal number of observations per bin. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

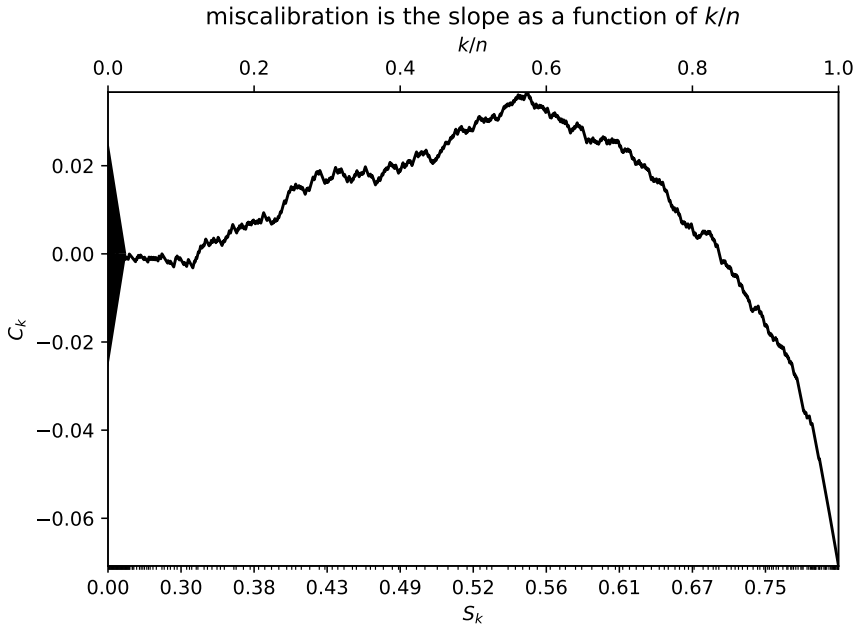

Figure 18: Cumulative plot for the sidewinder or horned rattlesnake (*Crotalus cerastes*), with sample size n = 1,300. The ECCE-MAD is $0.07081/\sigma_n = 5.446$, and the ECCE-R is $0.1075/\sigma_n = 8.267$; the associated asymptotic P-values are 1.0E–7 for the ECCE-MAD and zero to double-precision accuracy for the ECCE-R.

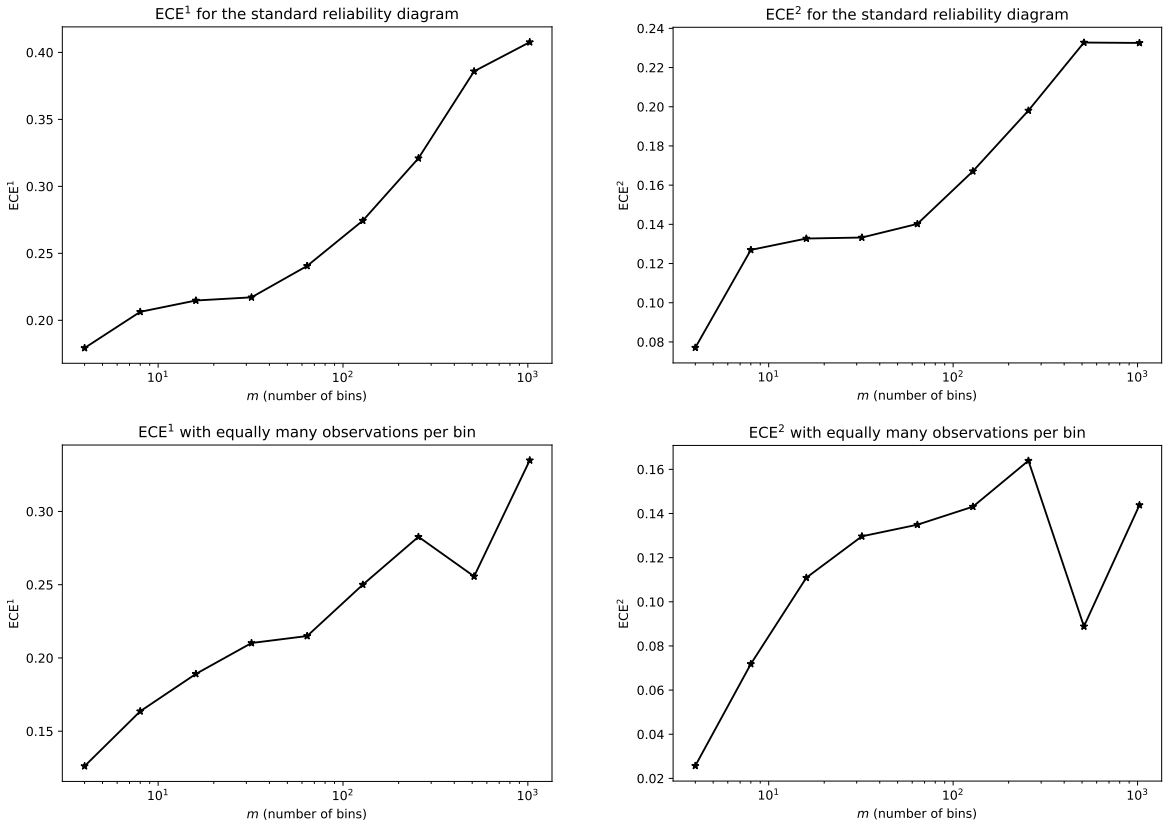

Figure 19: Empirical calibration errors for the Eskimo dog or husky, with sample size $n = 1{,}300$.

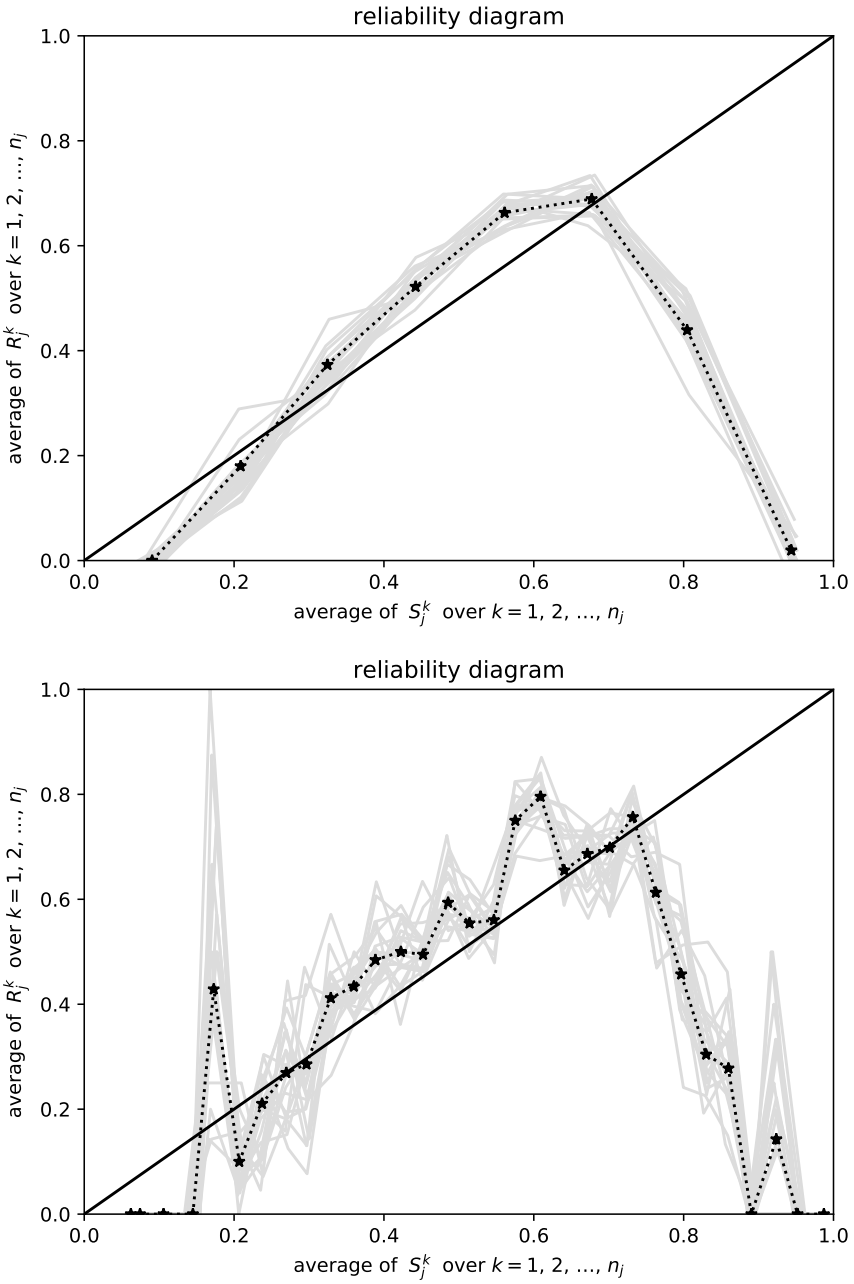

Figure 20: Reliability diagrams for the Eskimo dog or husky, with the bins roughly equispaced. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

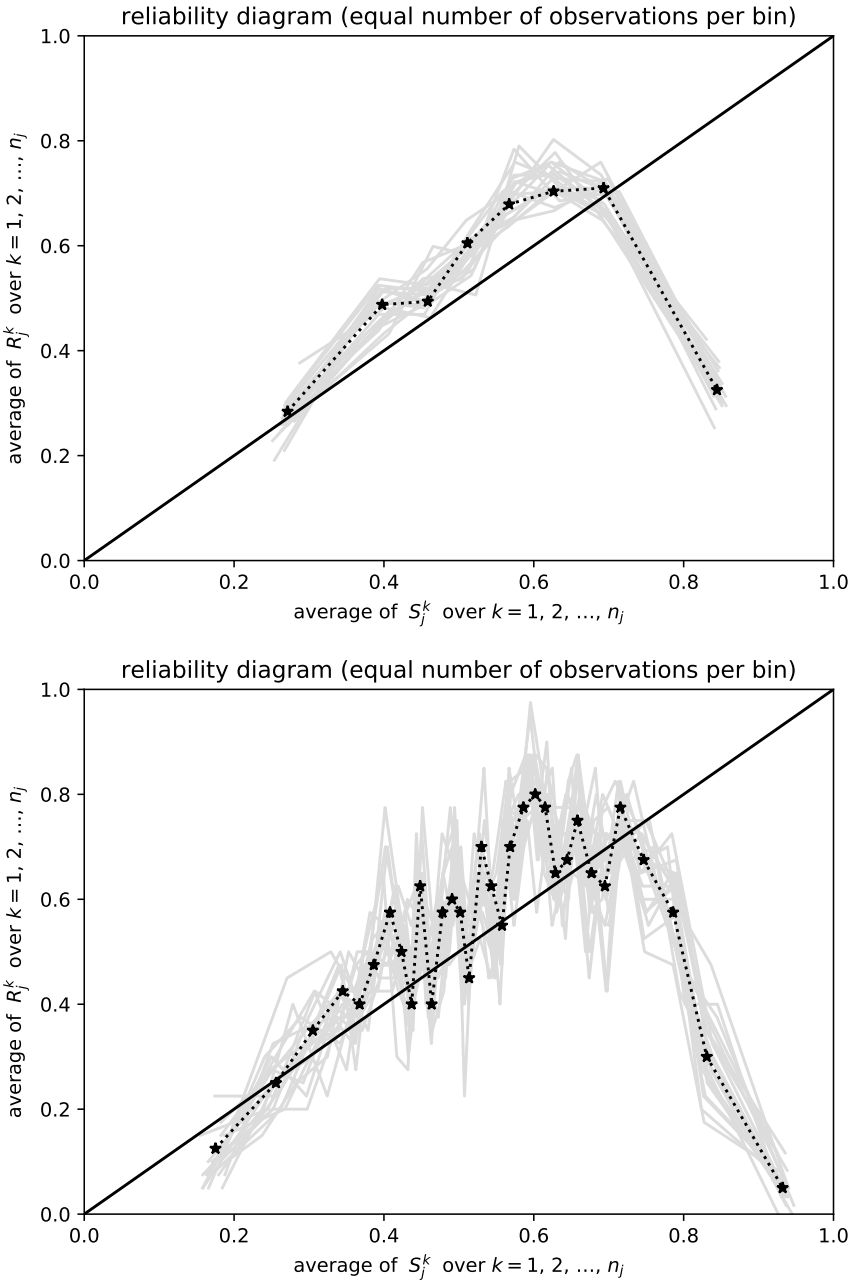

Figure 21: Reliability diagrams for the Eskimo dog or husky, with an equal number of observations per bin. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

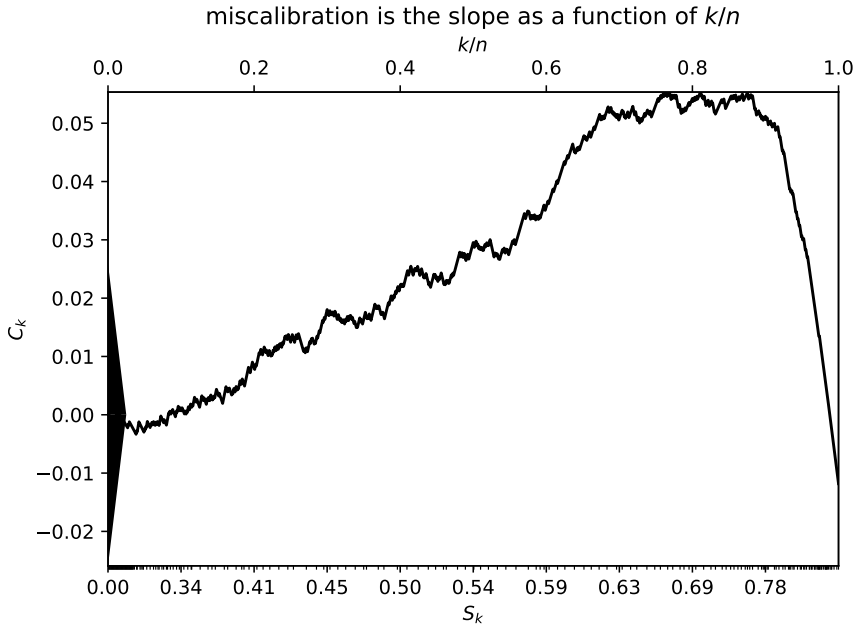

Figure 22: Cumulative plot for the Eskimo dog or husky, with sample size $n = 1{,}300$. The ECCE-MAD is $0.05534/\sigma_n = 4.274$, and the ECCE-R is $0.06715/\sigma_n = 5.186$; the associated asymptotic P-values are 3.8E–5 and 8.6E–7, respectively.

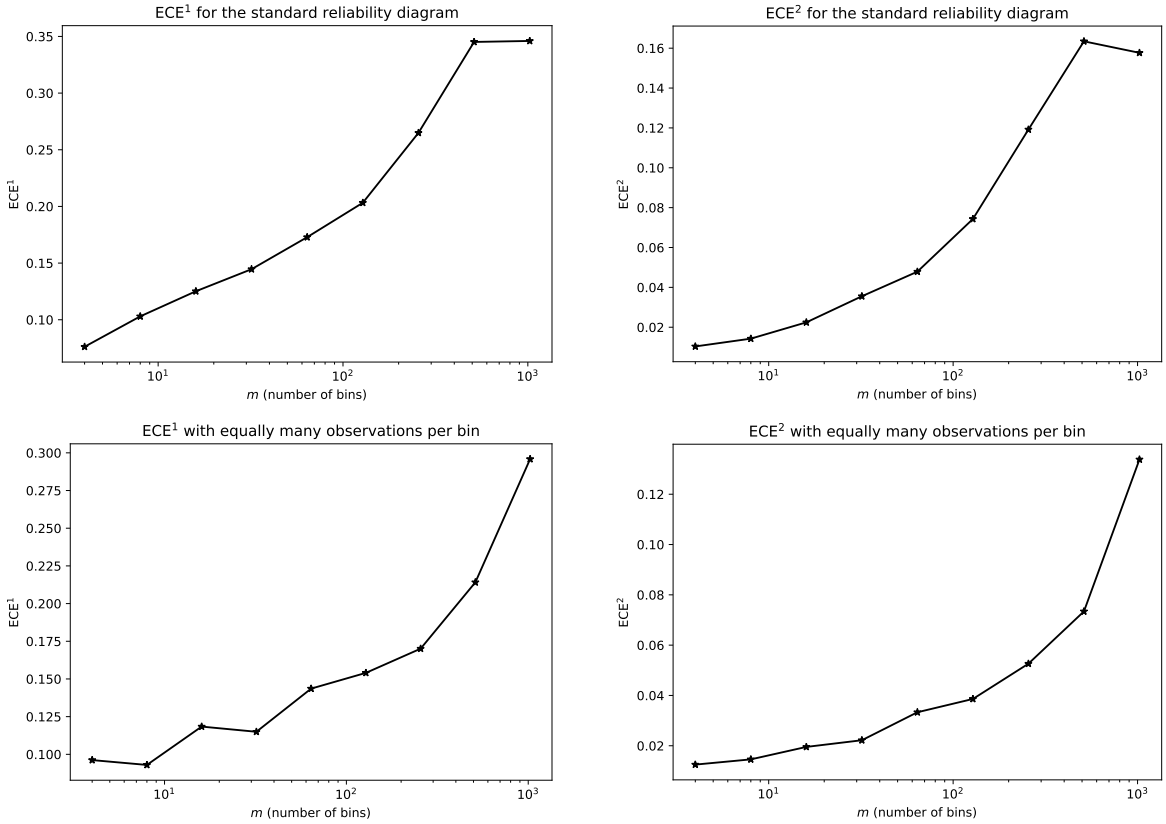

Figure 23: Empirical calibration errors for the wild boar (*Sus scrofa*), with sample size $n = 1{,}300$.

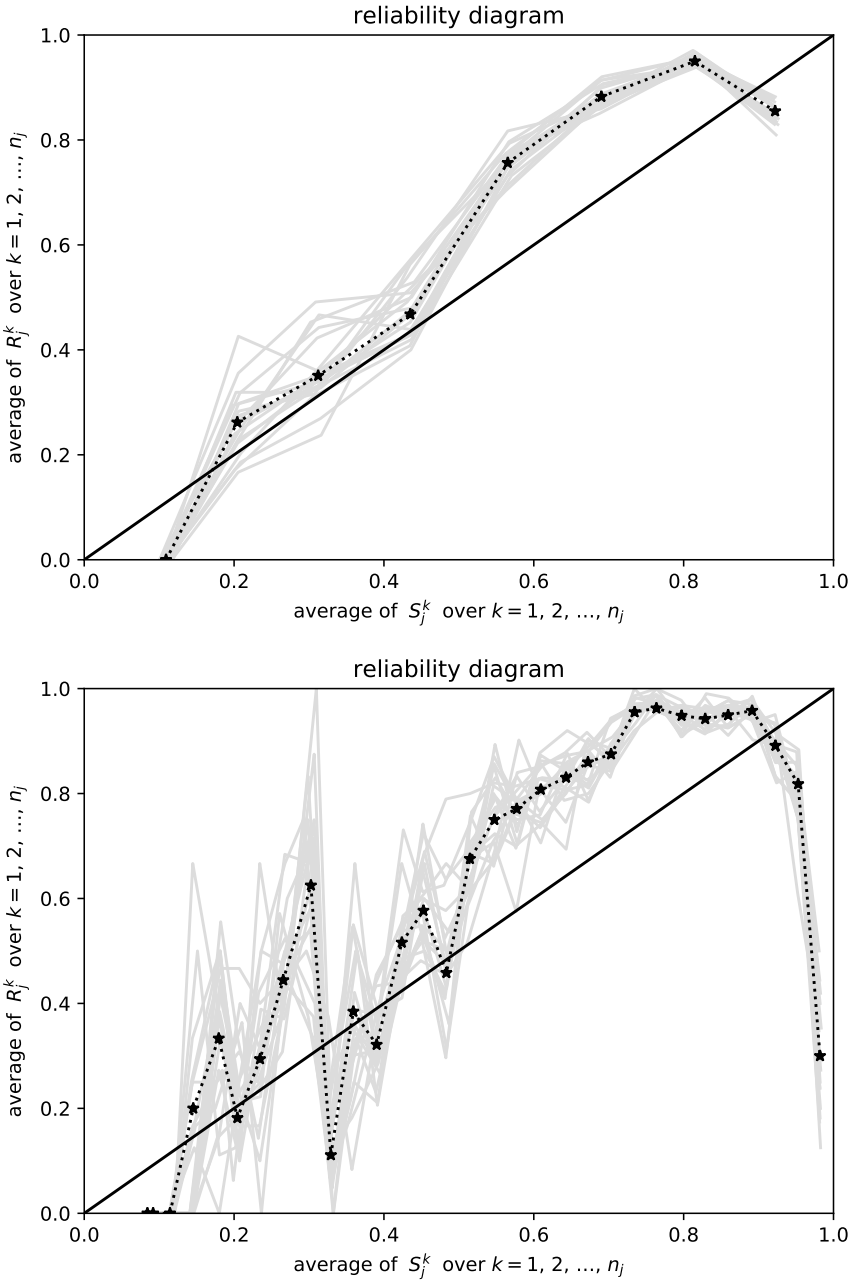

Figure 24: Reliability diagrams for the wild boar (*Sus scrofa*), with the bins roughly equispaced. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

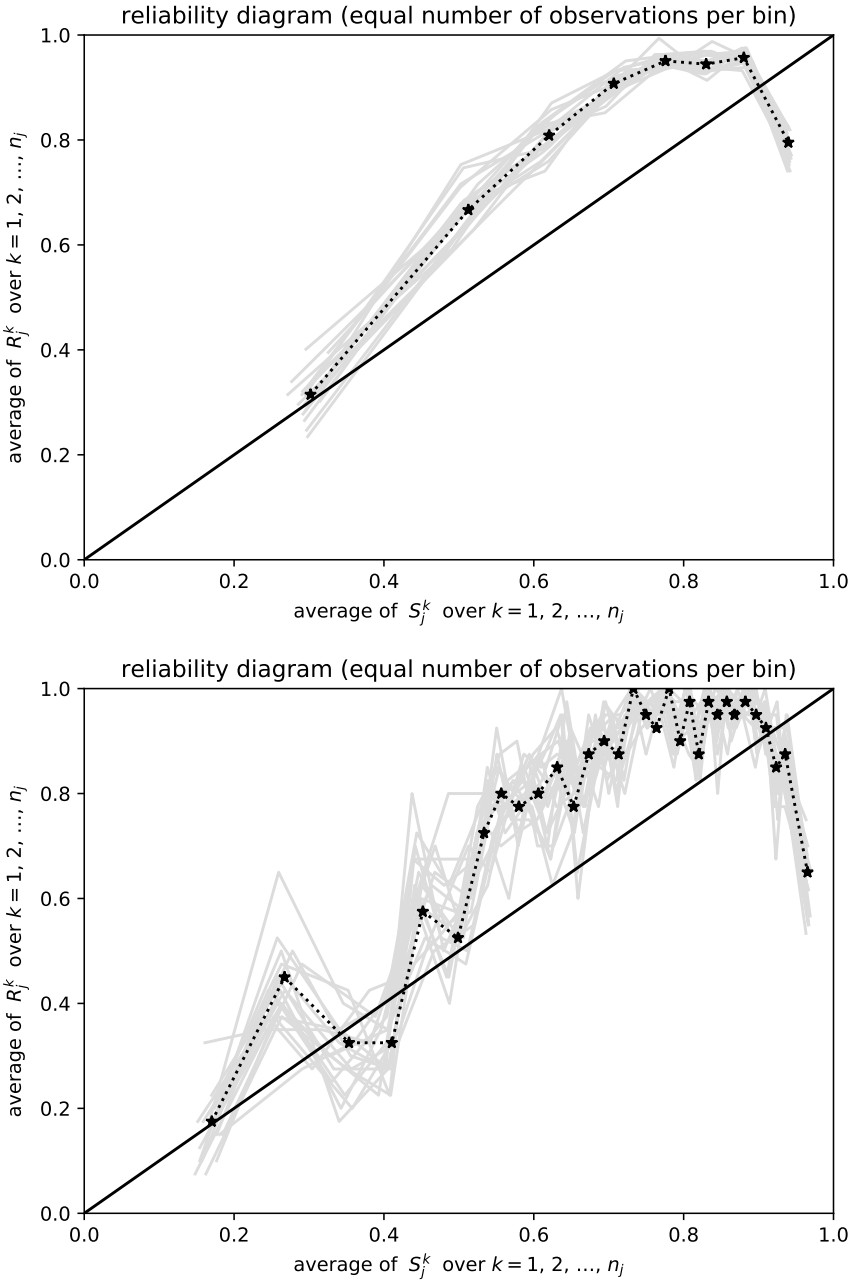

Figure 25: Reliability diagrams for the wild boar (*Sus scrofa*), with an equal number of observations per bin. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

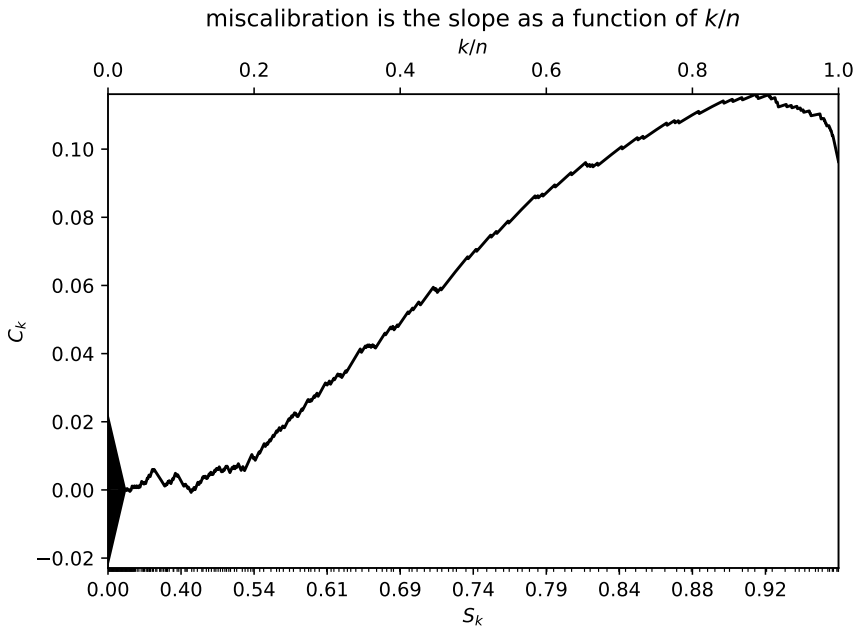

Figure 26: Cumulative plot for the wild boar (*Sus scrofa*), with sample size $n = 1{,}300$. The ECCE-MAD is $0.1161/\sigma_n = 10.14$, and the ECCE-R is $0.1172/\sigma_n = 10.23$; both associated asymptotic P-values are zero to double-precision accuracy.

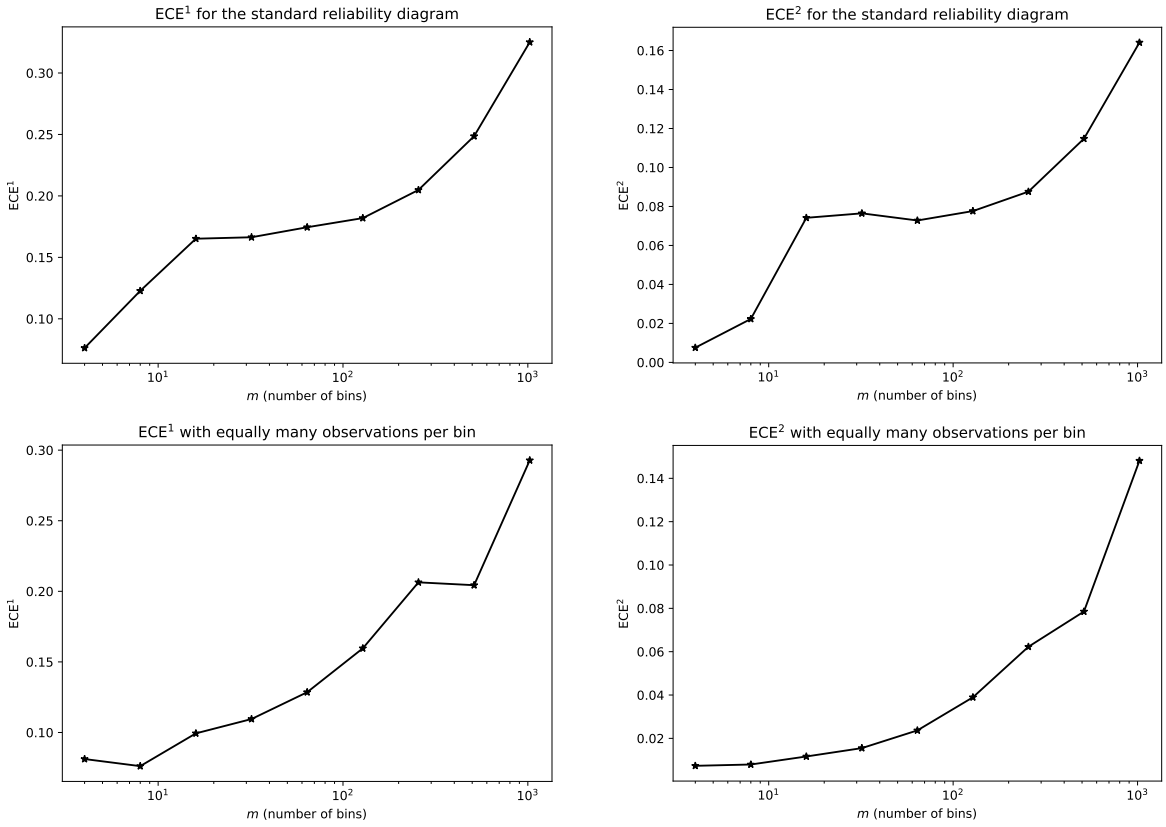

Figure 27: Empirical calibration errors for sunglasses, with sample size $n = 1{,}300$.

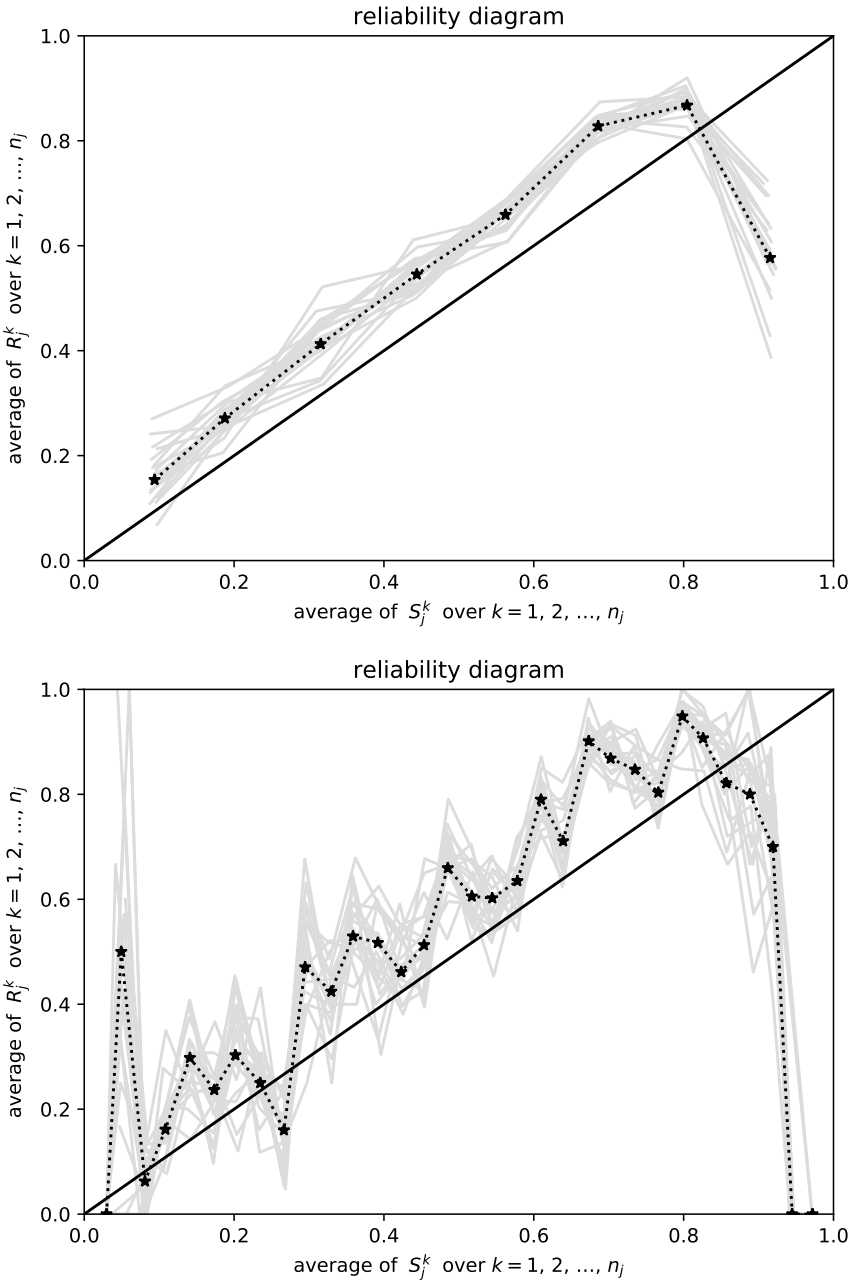

Figure 28: Reliability diagrams for sunglasses, with the bins roughly equispaced. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

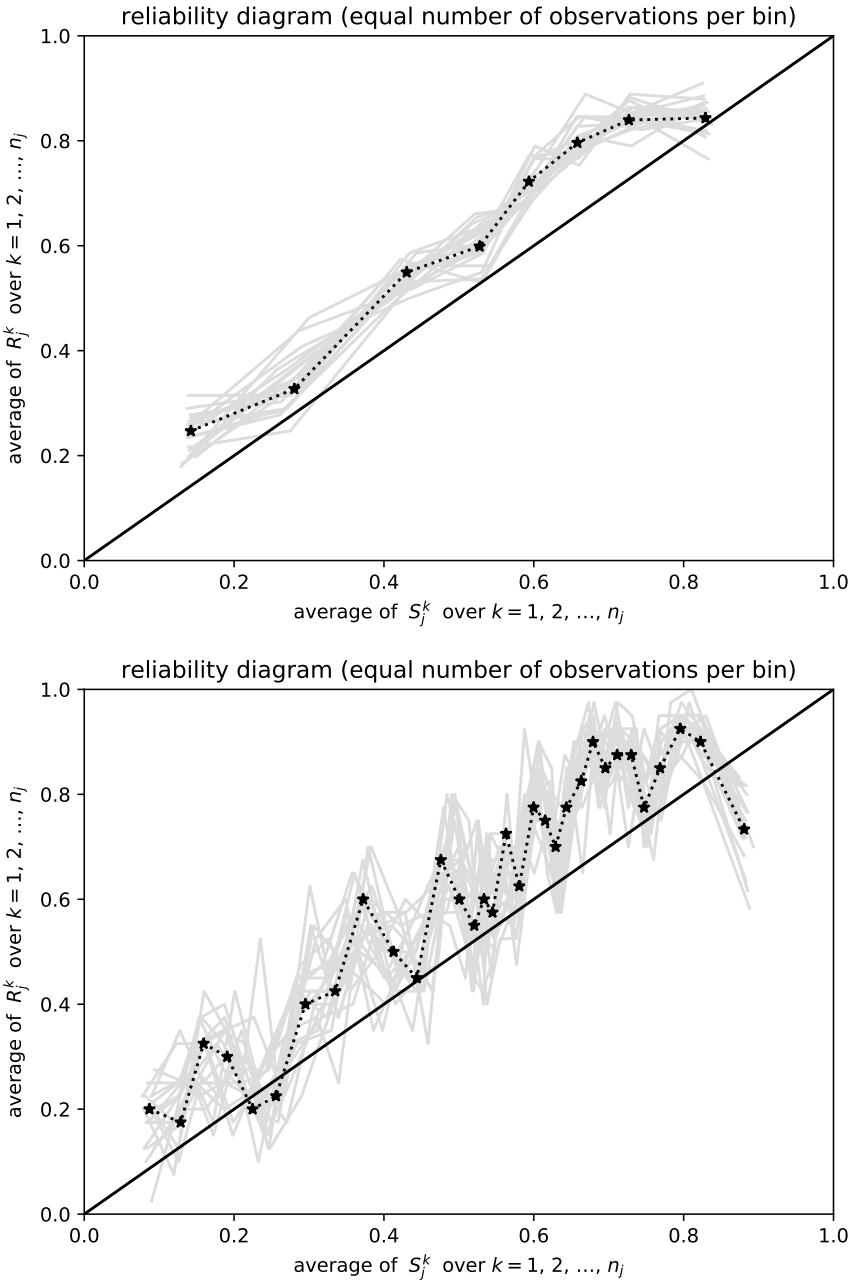

Figure 29: Reliability diagrams for sunglasses, with an equal number of observations per bin. There are $m = 8$ bins in the upper plot and $m = 32$ in the lower plot.

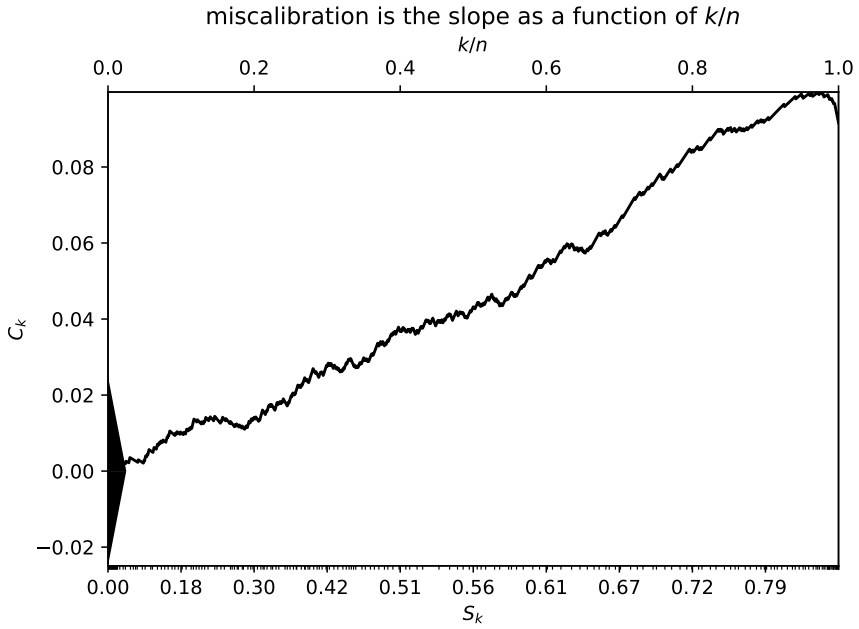

Figure 30: Cumulative plot for sunglasses, with sample size $n = 1{,}300$. The ECCE-MAD is $0.09972/\sigma_n = 8.004$, and the ECCE-R is $0.09977/\sigma_n = 8.008$; both associated asymptotic P-values are zero to double-precision accuracy.

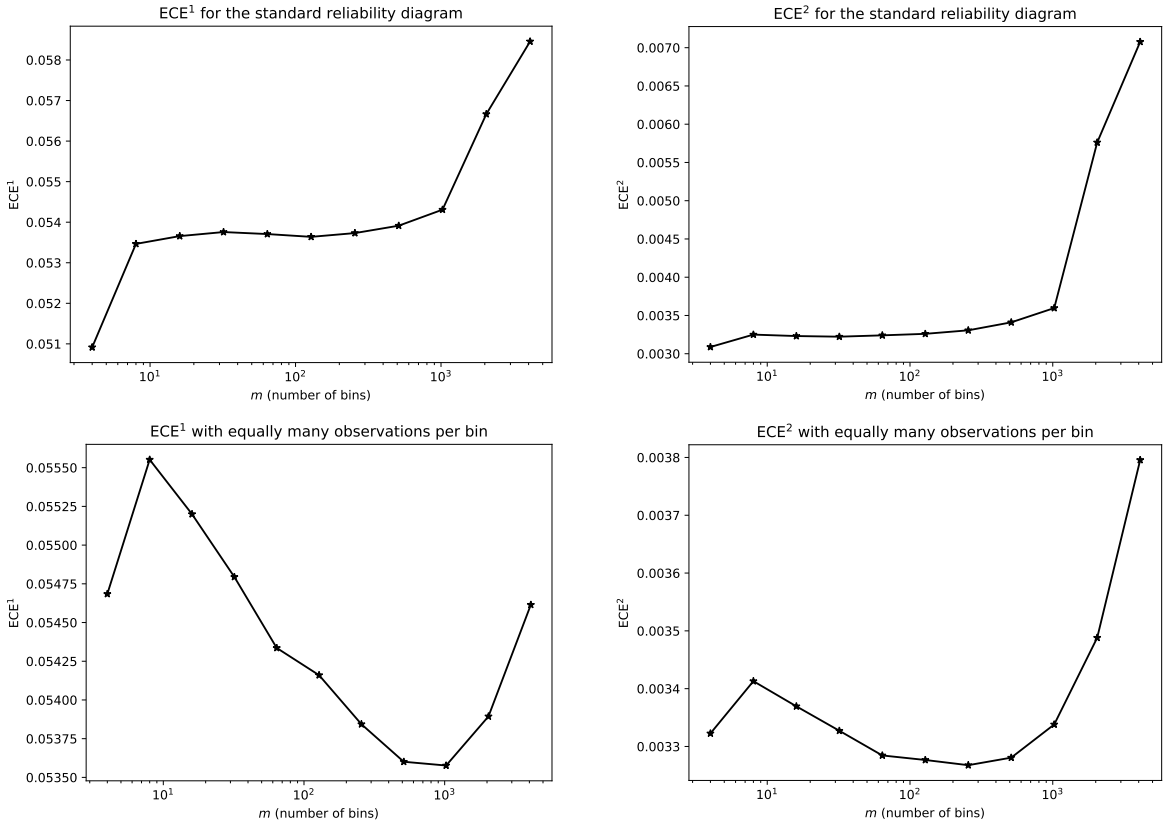

Figure 31: Empirical calibration errors for the full ImageNet-1000 training data set, with sample size $n = 1{,}281{,}167$.

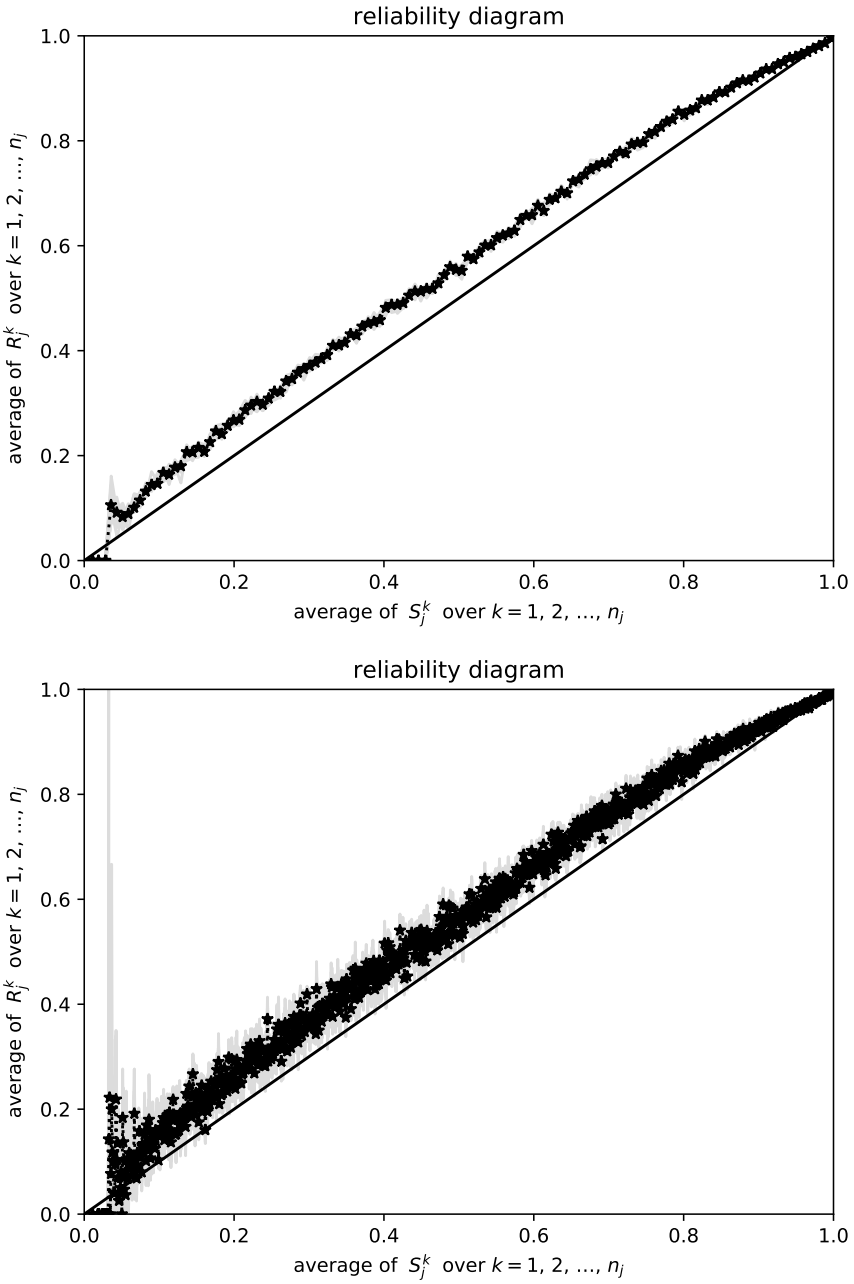

Figure 32: Reliability diagrams for the full ImageNet-1000 training data set, with the bins roughly equispaced. There are $m = 128$ bins in the upper plot and $m = 1{,}024$ in the lower plot.

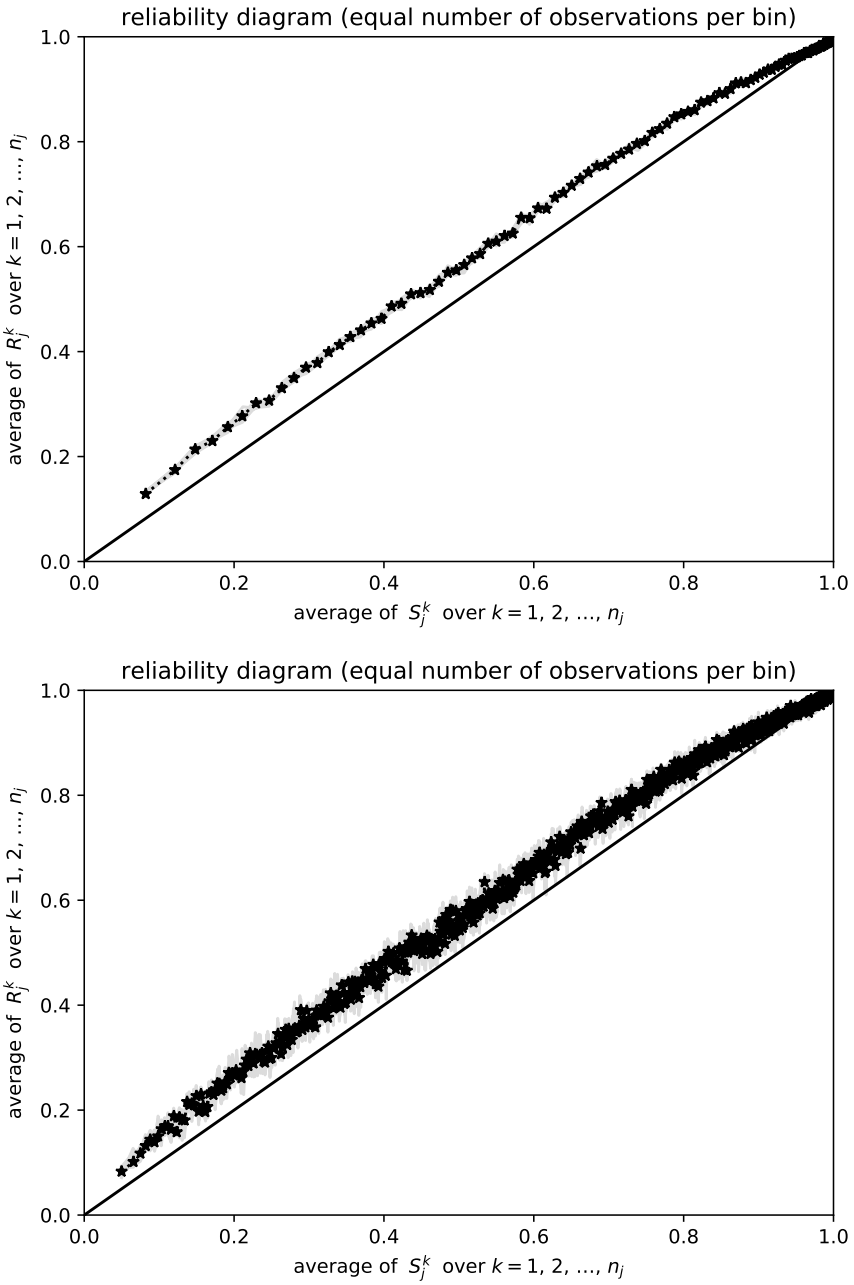

Figure 33: Reliability diagrams for the full ImageNet-1000 training data set, with an equal number of observations per bin. There are $m = 128$ bins in the upper plot and $m = 1{,}024$ in the lower plot.

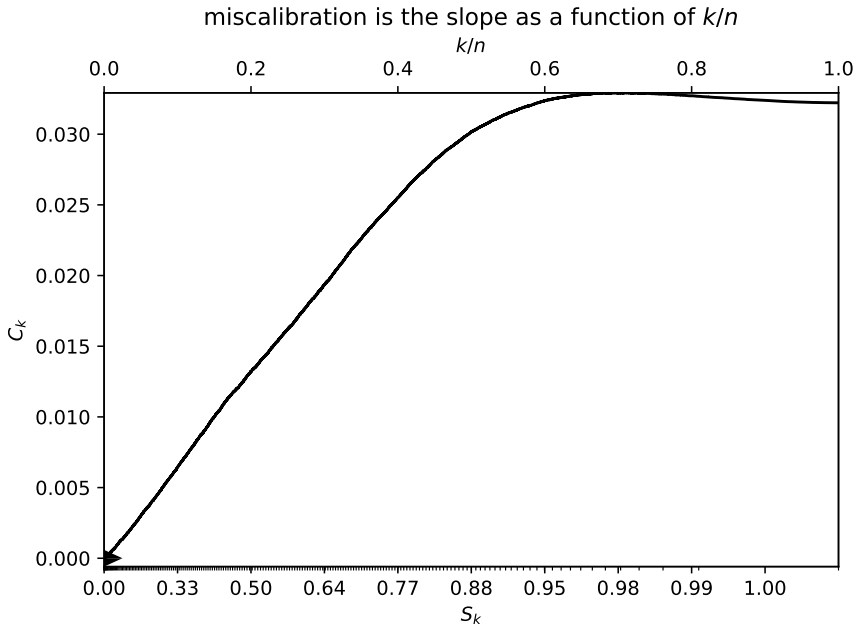

Figure 34: Cumulative plot for the full ImageNet-1000 training data set, with sample size n = 1,281,167. The ECCE-MAD is $0.03306/\sigma_n = 111.7$, and the ECCE-R is also $0.03306/\sigma_n = 111.7$; these indicate profoundly statistically significant miscalibration, courtesy of the large number of observations (the actual effect size is more modest, as seen by the values without dividing by σ_n). Both associated asymptotic P-values are zero to double-precision accuracy.

Appendix A. Proofs

This appendix provides rigorous proofs for the statements made in earlier sections.

The following is the proof of Theorem 1.

Proof (of Theorem 1): Since the responses are all independent, the expected value of the ECE^2 from (3) under the assumption of the null hypothesis is

$$\sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \sum_{k=1}^{n_j} \frac{S_j^k \left(1 - S_j^k\right)}{(n_j)^2}$$
$$= \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\frac{S_j^1 \left(1 - S_j^1\right)}{n_j} + \sum_{k=1}^{n_j} \frac{S_j^k \left(1 - S_j^k\right) - S_j^1 \left(1 - S_j^1\right)}{(n_j)^2} \right), \quad (16)$$

which is a Riemann sum for which the second sum in the right-hand side of (16) converges uniformly to 0 as $\max_{0 \le j \le m}(S_{j+1}^1 - S_j^1)$ tends to 0 (uniformly over j and independent of the values for n_j). The lower limit of the right-hand side of (16) converges to (8). ■

The following is the proof of Theorem 3.

Proof (of Theorem 3): It follows from the fact that both the scores and the responses fall on the unit interval $[0, 1]$ that

$$\left| \sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j} \right| \le \sum_{k=1}^{n_j} \frac{|R_j^k - S_j^k|}{n_j} \le \sum_{k=1}^{n_j} \frac{1}{n_j} = 1, \quad (17)$$

so

$$\left| \sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j} \right|^2 \le \left| \sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j} \right|. \quad (18)$$

It follows from (18) that

$$\sum_{j=1}^{m} (S_{j+1}^1 - S_j^1) \left| \sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j} \right| \ge \sum_{j=1}^{m} (S_{j+1}^1 - S_j^1) \left(\sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j} \right)^2, \quad (19)$$

which is the statement of the theorem expressed in terms of the definitions in (3) and (4). ■

The following is the proof of Theorem 8.

Proof (of Theorem 8): Since the responses are all independent, the expected value of the ECE^2 from (3) is

$$\sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \mathbb{E} \left[\left(\sum_{k=1}^{n_j} \frac{R_j^k - r(S_j^k) + r(S_j^k) - S_j^k}{n_j} \right)^2 \right]$$
$$= \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\left(\tilde{r}_j - \tilde{S}_j\right)^2 + \sum_{k=1}^{n_j} \frac{\mathbb{E}\left[\left(R_j^k - r(S_j^k) \right)^2 \right]}{(n_j)^2} \right), \quad (20)$$

where $\tilde{r}_j$ and $\tilde{S}_j$ denote the averages

$$\tilde{r}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} r(S_j^k) \tag{21}$$

and

$$\tilde{S}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} S_j^k \tag{22}$$

for $j = 1, 2, \ldots, m$. The fact that the variance of a Bernoulli distribution whose expected value is $r(S_j^k)$ is $r(S_j^k)(1 - r(S_j^k))$ yields

$$\sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\left(\tilde{r}_j - \tilde{S}_j\right)^2 + \sum_{k=1}^{n_j} \frac{\mathbb{E}\left[\left(R_j^k - r(S_j^k)\right)^2\right]}{(n_j)^2} \right)$$
$$= \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\left(\tilde{r}_j - \tilde{S}_j\right)^2 + \sum_{k=1}^{n_j} \frac{r(S_j^k)\left(1 - r(S_j^k)\right)}{(n_j)^2} \right). \tag{23}$$

Referencing terms in each bin to the same score yields

$$\sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\left(\tilde{r}_j - \tilde{S}_j\right)^2 + \sum_{k=1}^{n_j} \frac{r(S_j^k)\left(1 - r(S_j^k)\right)}{(n_j)^2} \right)$$
$$= \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\left(r(S_j^1) - S_j^1\right)^2 + \frac{r(S_j^1)\left(1 - r(S_j^1)\right)}{n_j} \right)$$
$$+ \sum_{j=1}^{m} \left(S_{j+1}^1 - S_j^1\right) \left(\left(\tilde{r}_j - \tilde{S}_j\right)^2 - \left(r(S_j^1) - S_j^1\right)^2 + \sum_{k=1}^{n_j} \frac{r(S_j^k)\left(1 - r(S_j^k)\right) - r(S_j^1)\left(1 - r(S_j^1)\right)}{(n_j)^2} \right), \tag{24}$$

which is the sum of two Riemann sums, the latter of which converges to 0 as $\max_{0 \le j \le m}(S_{j+1}^1 - S_j^1)$ tends to 0. The lower limit of the right-hand side of (24) converges to (10), so combining (20)–(24) completes the proof. ■

The following is the proof of Theorem 11.

Proof (of Theorem 11): Applying first {1} the Chernoff or Hoeffding bound for averages of independent Bernoulli variates to deviate from their expected values by more than $n^{-1/4}$ (or for the unnormalized sums to deviate from their expected values by more than $n^{3/4}$), then {2} union bounding across the averages for $k = 1, 2, \ldots, n$, and finally {3} applying the Borel-Cantelli Lemma over the sample size n yields that as n becomes arbitrarily large the ECCE-MAD defined in (6) converges almost surely to

$$\lim_{n \to \infty} \max_{1 \le k \le n} \left| \frac{1}{n} \sum_{j=1}^{k} \left(r(S_j) - S_j\right) \right| = \max_{0 \le t \le 1} \left| \int_0^t \left(r(s) - s\right) dF(s) \right| \tag{25}$$

and that the ECCE-R defined in (7) converges almost surely to

$$\lim_{n\to\infty} \left(\max_{0\le k\le n} \frac{1}{n} \sum_{j=0}^{k} (r(S_j) - S_j) - \min_{0\le k\le n} \frac{1}{n} \sum_{j=0}^{k} (r(S_j) - S_j) \right) = \max_{0\le t\le 1} \int_0^t (r(s) - s)\, dF(s) - \min_{0\le t\le 1} \int_0^t (r(s) - s)\, dF(s). \quad (26)$$

The dominated convergence theorem then yields convergence of the expected values of the ECCE-MAD and the ECCE-R to the same values in (25) and (26), courtesy of the domination

$$\left| \frac{1}{n} \sum_{j=1}^{k} (R_j - S_j) \right| \le \frac{1}{n} \sum_{j=1}^{n} |R_j - S_j| \le 1 \quad (27)$$

for $k = 1, 2, \ldots, n$ (recall that both R_j and S_j lie in the unit interval $[0, 1]$, so $|R_j - S_j| \le 1$).

Now, if (25) were 0, then

$$\int_0^t (r(s) - s)\, dF(s) = 0 \quad (28)$$

for all t in the unit interval $[0, 1]$; differentiating with respect to t would then show that $r(s) = s$ except on a set of measure 0 relative to dF, making $\int_0^1 |r(s) - s|\, dF(s)$ be 0, too. The assumption that $\int_0^1 |r(s) - s|\, dF(s) > 0$ thus implies that the limit (25) of the expected value of the ECCE-MAD must be strictly positive, completing the proof for the ECCE-MAD.

Similarly, if (26) were 0, then

$$\int_0^t (r(s) - s)\, dF(s) = c \quad (29)$$

for all t in the unit interval $[0, 1]$, for some real number c (after all, the maximum and minimum of a function are the same only if the function is equal to some constant c); as before, differentiating both sides of (29) with respect to t would then show that $r(s) = s$ except on a set of measure 0 relative to dF, making $\int_0^1 |r(s) - s|\, dF(s)$ be 0, too. The assumption that $\int_0^1 |r(s) - s|\, dF(s) > 0$ thus implies that the limit (26) of the expected value of the ECCE-R must be strictly positive, completing the proof for the ECCE-R. ■

Appendix B. Additional figures

This appendix provides analogues for the samples sizes n = 8,192 and n = 131,072 of Figure 1 from Section 3.1 (Figure 1 corresponds to the sample size n = 32,768). Figure 35 is for n = 8,192; Figure 36 is for n = 131,072.

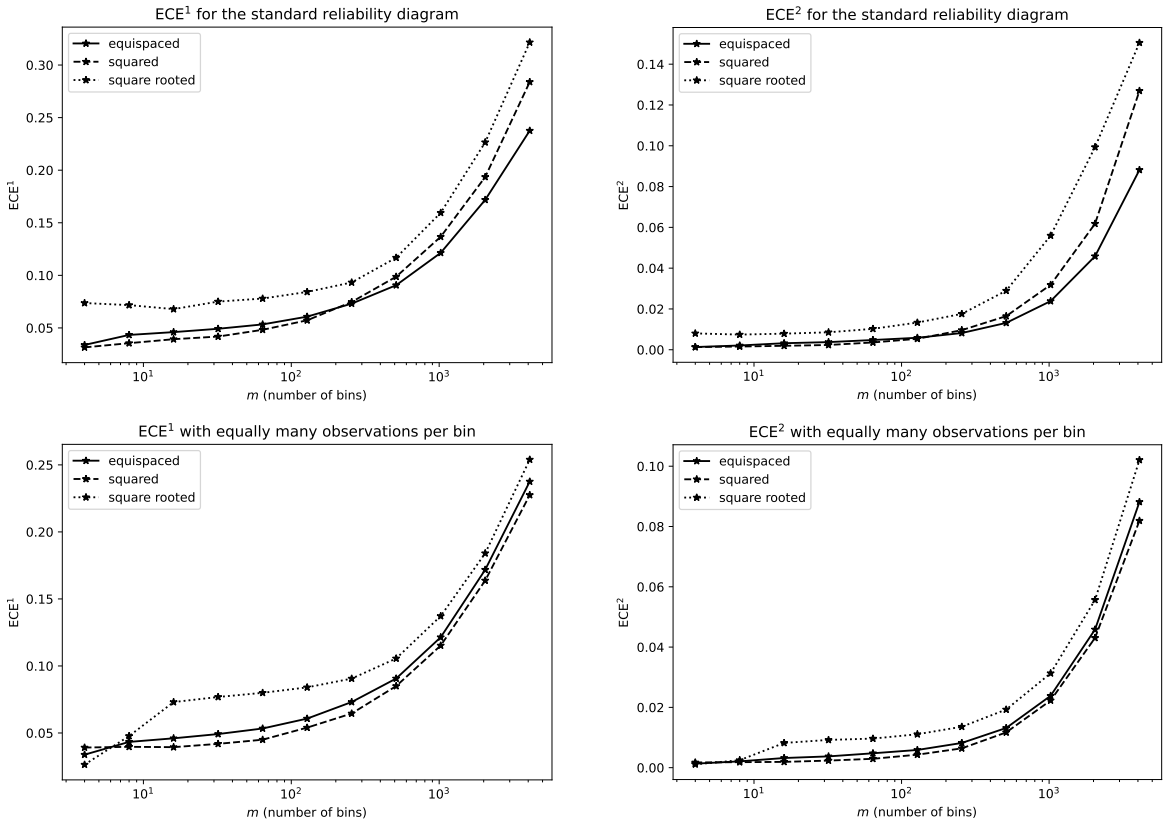

Figure 35: Empirical calibration errors for the synthetic data set with the sample size $n = 8{,}192$; the scores are equispaced, squared, or square rooted, as indicated in the legends.

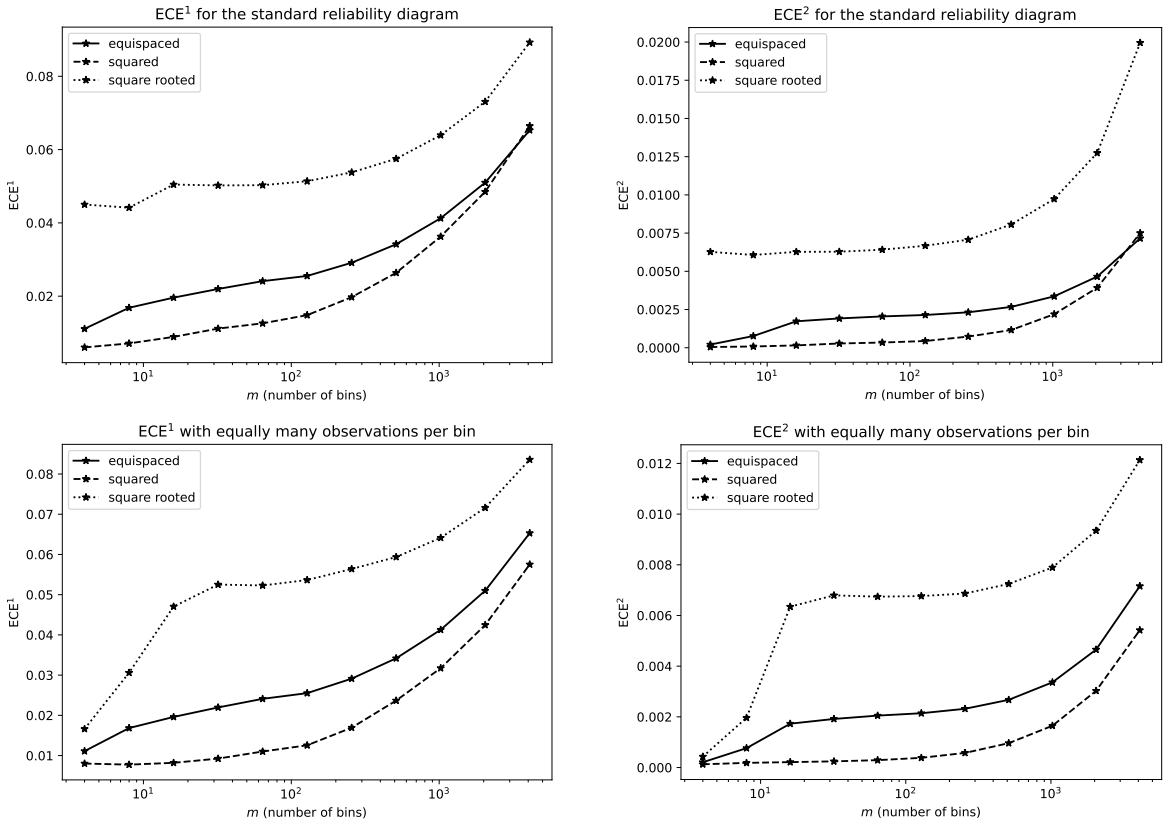

Figure 36: Empirical calibration errors for the synthetic data set with the sample size $n = 131{,}072$; the scores are equispaced, squared, or square rooted, as indicated in the legends.

References

Jochen Bröcker. Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review*, 136(11):4488–4502, 2008.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. Technical Report 2006.12800, arXiv, 2020. URL `https://arxiv.org/abs/2006.12800`. Also published as a poster and paper at the 2021 International Conference on Learned Representations (ICLR).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

Andrei N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione (On the empirical determination of a distribution function). *Giornale dell'Istituto Italiano degli Attuari*, 4: 83–91, 1933.

Nicolaas H. Kuiper. Tests concerning random points on a circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen Series A*, 63:38–47, 1962.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems 32*, pages 3792–3803. Curran Associates, 2019.

Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-Cal: an optimal test for the calibration of predictive models. Technical Report 2203.01850, arXiv, 2022. URL `https://arxiv.org/abs/2203.01850`.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the 2019 Computer Vision and Pattern Recognition Workshops*, pages 38–41. Computer Vision Foundation, IEEE, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, 2019.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. Technical Report 2012.08668, arXiv, 2020. URL `https://arxiv.org/abs/2012.08668`. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022)*, forthcoming.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Jeffrey S. Simonoff and Frederic Udina. Measuring the stability of histogram appearance when the anchor position is changed. *Computational Statistics and Data Analysis*, 23(3):335–353, 1997.

Nikolai Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Université de Moscou*, 2(2):3–11, 1939.

Ramidha Srihera and Winfried Stute. Nonparametric comparison of regression functions. *Journal of Multivariate Analysis*, 101(9):2039–2059, 2010.

Winfried Stute. Nonparametric model checks for regression. *Annals of Statistics*, 25(2):613–641, 1997.

Mark Tygert. Cumulative deviation of a subpopulation from the full population. *Journal of Big Data*, 8(117):1–60, 2021a. URL `https://arxiv.org/abs/2008.01779`.

Mark Tygert. A graphical method of cumulative differences between two subpopulations. *Journal of Big Data*, 8(158):1–29, 2021b. URL `https://arxiv.org/abs/2108.02666`.

Mark Tygert. Calibration of P-values for calibration and for deviation of a subpopulation from the full population. Technical Report 2202.00100, arXiv, 2022. URL `https://arxiv.org/abs/2202.00100`.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. Evaluating model calibration in classification. *Proceedings of Machine Learning Research*, 89:3459–3467, 2019. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*.