Distribution-free binary classification: prediction sets, confidence intervals and calibration

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas

Kiho Park
parkkiho@uchicago.edu

1 Introduction

Consider a predictor $f : \mathcal{X} \to \mathcal{Z}$ with $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$ for binary classification. If $\mathcal{Z} = [0, 1]$, it is tempting to interpret $f(X)$ as a probability, and hope that $f(x) \approx \mathbb{P}(Y = 1 \mid X) = \mathbb{E}[Y \mid X]$. However, such hope is unfounded. This paper discusses three notions of uncertainty quantification: calibration, prediction sets (PS), and confidence intervals (CI). A function $f : \mathcal{X} \to [0, 1]$ is said to be (perfectly) calibrated if

$$\mathbb{E}[Y \mid f(X) = a] = a \quad \text{a.s. for all } a \text{ in the range of } f$$

Define the set of all subsets of $\mathcal{Y}$, $\mathcal{L} := \{\{0\}, \{1\}, \{0, 1\}, \varnothing\}$, and fix $\alpha \in (0, 1)$. A function $S : \mathcal{X} \to \mathcal{L}$ is a $(1 - \alpha)$-PS if

$$P(Y \in S(X)) \geq 1 - \alpha$$

Finally, let $\mathcal{I}$ denote the set of all subintervals of $[0, 1]$. A function $C : \mathcal{X} \to \mathcal{I}$ is a $(1 - \alpha)$-CI if

$$P(\mathbb{E}[Y \mid X] \in C(X)) \geq 1 - \alpha$$

Most of our results are in the distribution-free setting. This paper is based on the statistical setup of post-hoc uncertainty quantification, which is a common paradigm where the available labeled data is split into a training set and a calibration set. All results in this paper are conditional on the training set. We denote the calibration set as $\mathcal{D}_n = \{(X_i, Y_i)\}_{i \in [n]}$. A prototypical test point is denoted as (X_{n+1}, Y_{n+1}). The learner observes realized values of all random variables (X_i, Y_i), except Y_{n+1}.

2 Calibration, confidence intervals and prediction sets

2.1 Approximate and asymptotic calibration

Proposition 1. *A function $f : \mathcal{X} \to [0, 1]$ is perfectly calibrated if and only if there exists a space $\mathcal{Z}$ and a function $g : \mathcal{X} \to \mathcal{Z}$, such that*

$$f(x) = \mathbb{E}[Y \mid g(X) = g(x)] \quad \textit{almost surely } P_X$$

Since Proposition 1 states that calibration is possible iff the RHS is known exactly for some g, perfect calibration is impossible in practice. Thus we resort to satisfying the requirement approximately, which is implicitly the goal of many empirical calibration techniques.

Definition 1 (Approximate calibration)**.** *A predictor $f : \mathcal{X} \to [0, 1]$ is (ε, α)-calibrated for some $\varepsilon, \alpha \in [0, 1]$ if with probability at least $1 - \alpha$,*

$$|\mathbb{E}[Y \mid f(X)] - f(X)| \leq \varepsilon$$

We are typically only interested in the smallest "pareto optimal boundary" pairs of (ε, α) for which approximate calibration holds, or a specifically for a fixed α.

Suppose f is not approximately calibrated for small values of ε and α. We can recalibrate f using a post-hoc calibration algorithm $\mathcal{A}$ which outputs $\mathcal{A}(\mathcal{D}_n, f) = h_n : \mathcal{X} \to [0,1]$. In this paper, we study when $\mathcal{A}$ can be shown to satisfy distribution-free approximate calibration:

$$P^{n+1}(|\mathbb{E}[Y \mid h_n(X_{n+1})] - h_n(X_{n+1})| \leq \varepsilon) \geq 1 - \alpha \text{ for every } f, P \tag{1}$$

where P^{n+1} denotes the product distribution of the i.i.d. calibration and test points.

We use $(\mathcal{X} \times \mathcal{Y})^* = \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n$ to denote the space of the calibration data for arbitrary n, and $[0,1]^{\mathcal{X}}$ to denote the space of a function from $\mathcal{X}$ to $[0,1]$.

Definition 2 (Distribution-free asymptotic calibration)**.** *A post-hoc calibration algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \times [0,1]^{\mathcal{X}} \to [0,1]^{\mathcal{X}}$ is said to be distribution-free asymptotically calibrated if there exists an $\alpha \in (0, 0.5)$ and a $[0,1]$-valued sequence $\{\varepsilon_n\}_{n \in \mathbb{N}}$ with $\lim_{n \to \infty} \varepsilon_n = 0$, such that for every n, $h_n = \mathcal{A}(\mathcal{D}_n, f)$ satisfies condition (1) with parameters (ε_n, α).*

2.2 Prediction sets and confidence interval with respect to f

Barber (2020) show that if an algorithm provides $(1-\alpha)$-CI for all product distributions P^{n+1}, then it also provides a $(1-\alpha)$-PS whenever the distribution of P_X is nonatomic. It implies that distribution-free CIs must necessarily be wide. This can be treated as an impossibility result for the existence of informative distribution-free CIs. One way to circumvent it is to consider CIs at a 'coarser resolution' as follows.

Definition 3 (CI or PS w.r.t. f)**.** *Fix a predictor $f : \mathcal{X} \to [0,1]$ and let $(X, Y) \sim P$. A function $C : [0,1] \to \mathcal{I}$ is said to be a $(1-\alpha)$-CI with respect to f if*

$$P(\mathbb{E}[Y \mid f(X)] \in C(f(X))) \geq 1 - \alpha$$

Analogously, a function $S : [0,1] \to \mathcal{L}$ is a $(1-\alpha)$-PS with respect to f if

$$P(Y \in S(f(X))) \geq 1 - \alpha$$

Let $\mathcal{C}$ denote an algorithm that produces a CI function using f and $\mathcal{D}_n$, $C_n = \mathcal{C}(\mathcal{D}_n, f) : [0,1] \to \mathcal{I}$. Similarly, let $\mathcal{S}$ denote an algorithm that produces a PS function, $S_n = \mathcal{S}(\mathcal{D}_n, f) : [0,1] \to \mathcal{L}$. C_n is said to be a distribution-free CI w.r.t. a fixed f if

$$P^{n+1}(\mathbb{E}[Y_{n+1} \mid f(X_{n+1})] \in C_n(f(X_{n+1}))) \geq 1 - \alpha \quad \text{for every } P$$

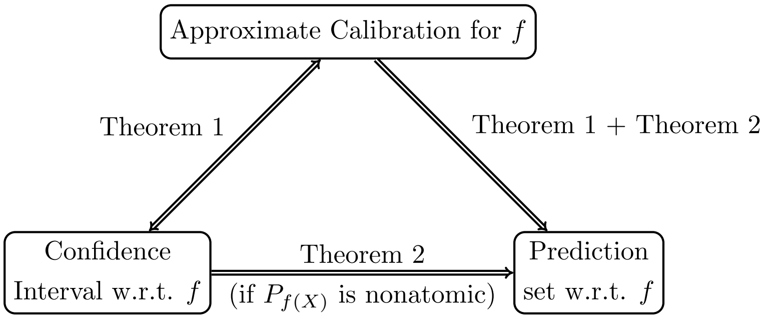
and S_n is said to be a distribution-free PS w.r.t. a fixed f if

$$P^{n+1}(Y_{n+1} \in S_n(f(X_{n+1}))) \geq 1 - \alpha \quad \text{for every } P$$

2.3 When is distribution-free post-hoc uncertainty quantification possible?

Are distribution-free guarantees too restrictive, or can they be achieved? We show that the answer for calibration and CIs depends on how large the range of f is. We also show a close relationship between approximate calibration and CIs w.r.t. f. In contrast, it is well known that meaningful and informative distribution-free PSs can be produced for any f, using a split conformal prediction.

3 Relating the notions of uncertainty quantification


Approximate Calibration for f
Theorem 1
Theorem 1 + Theorem 2
Confidence Interval w.r.t. f
Theorem 2
(if $P_{f(X)}$ is nonatomic)
Prediction set w.r.t. f

3.1 Relating calibration and confidence intervals

Theorem 1. *Fix any $\alpha \in (0,1)$. Let $f : \mathcal{X} \to [0,1]$ be a predictor that is (ε, α)-calibrated for some $\varepsilon \in (0,1)$. Then, the function C such that $C(z) := [z-\varepsilon, z+\varepsilon] \cap [0,1]$ for $z \in [0,1]$ is a $(1-\alpha)$-CI with respect to f.*

Conversely, fix a scoring function $f : \mathcal{X} \to [0,1]$. If C is a $(1-\alpha)$-CI with respect to f, then the predictor $\widetilde{f}$, such that $\widetilde{f}(x) := m_C(f(x))$ for $x \in \mathcal{X}$ where $m_C(z) := (u_C(z) + l_C(z))/2$ for $u_C(z) := \sup\{t : t \in C(z)\}$ and $l_C(z) := \inf\{t : t \in C(z)\}$, is (ε, α)-calibrated for $\varepsilon = \sup_{z \in [0,1]}\{|C(z)|/2\}$ where $|C(\cdot)|$ denotes the width of CI.

Note that Theorem 1 is not restricted to the post-hoc uncertainty quantification setting and the calibration and CI functions need not satisfy distribution-free guarantees.

3.2 Relating confidence intervals and prediction sets in the distribution-free setting

For a fixed function $f : \mathcal{X} \to [0,1]$, define $\mathcal{P}_f := \{\text{distributions } P \text{ over } \mathcal{X} \times \mathcal{Y} : P_{f(X)} \text{ is nonatomic}\}$.

Theorem 2. *Fix $f : \mathcal{X} \to [0,1]$ and $\alpha \in (0,1)$. If C_n is a distribution-free confidence interval with respect to f, then $S_n(\cdot) = C_n(\cdot) \cap \{0,1\}$ is a $(1-\alpha)$-prediction set with respect to f for every $P \in \mathcal{P}_f$.*

Theorem 2 can be interpreted as an impossibility result for distribution-free valid CIs that are 'informative' for all distributions. If $\mathcal{P}_f$ is non-empty, then no distribution-free CI function can be distribution consistent for C_p where $C_p(\cdot) = \{\mathbb{E}_P[Y \mid f(X) = \cdot]\}$.

Corollary 1. *Fix $f : \mathcal{X} \to [0,1]$ and $\alpha \in (0, 0.5)$. If C_n is a distribution-free confidence interval with repect to f, and $\mathcal{P}_f$ is non-empty, then there exists a distribution $P \in \mathcal{P}_f$ such that*

$$\mathbb{E}_{P^{n+1}}|C_n(f(X_{n+1})| \geq 0.5 - \alpha$$

Corollary 1 show that if $\mathcal{P}_f$ is non-empty, then no distribution-free CI function can be 'distribution-free consistent for $C_p(\cdot) = \{\mathbb{E}_P[Y \mid f(X) = \cdot]\}$. Thus we would like to know when $\mathcal{P}_f$ is non-empty. If the range of f is countable, $\mathcal{P}_f$ is empty. On the other hand, if the range of f is $[0,1]$ or contains any sub-interval of $[0,1]$, then $\mathcal{P}_f$ is non-empty.

3.3 Impossibility result for distribution-free post-hoc calibration

Let $\mathcal{X}_z = \{x \in \mathcal{X}; g(x) = z\}$ for g in Proposition 1 and define $f_z = P(Y = 1 \mid X \in \mathcal{X}_z)$.

Corollary 2 (to Proposition 1). *Any calibrated classifier f is characterized by an index set $\mathcal{Z}$,*

(A) a partition of $\mathcal{X}$ into subsets $\{\mathcal{X}_z\}_{z\in\mathbb{Z}}$, and

(B) corresponding conditional probabilities $\{f_z\}_{z\in\mathcal{Z}}$.

Existing ML techniques often implicitly do (A). They produce f that, while miscalibrated, may have some rough monotonicity with respect to the true probability. A post-hoc calibration method 'recalibrates' f by mapping its output to a new value in $[0, 1]$. Let $m_n : [0, 1] \to [0, 1]$ be the implicit mapping function so that $h_n(x) = m_n(f(x))$ for $h_n = \mathcal{A}(\mathcal{D}_n, f)$. The mapping m_n learnt by some popular parametric algorithms is strictly monotonic, and hence, injective. We call these as 'injective' post-hoc calibration algorithms. We now state the impossibility result for distribution-free calibration.

Theorem 3. *It is impossible for an injective post-hoc calibration algorithm to be distrbition-free asymptotically calibrated.*

The implication of Theorem 3 is that injective algorithms cannot satisfy distribution-free calibration in any meaningful way. We conjecture that a result like Theorem 3 holds even more generally for any parametric post-hoc calibration method. Some nonparametric calibration methods are not injective, and thus can potentially satisfy distribution-free asymptotic calibration guarantees.

4 Achieving distribution-free calibration

4.1 Distribution-free calibration given a fixed sample-space partition

Suppose we have a fixed partition of $\mathcal{X}$ into regions $\{\mathcal{X}_b\}_{b\in[B]}$, and let $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$. Denote the partition-identity function as $\mathcal{B} : \mathcal{X} \to [B]$ where $\mathcal{B}(x) = b \iff x \in \mathcal{X}_b$. Given a calibration set $\{(X_i, Y_i)\}_{i\in[n]}$, let $N_b := |\{i \in [n] : \mathcal{B}(X_i) = b\}|$. In this subsection, we assume that $N_b \geq 1$. Define

$$\widehat{\pi}_b := \frac{1}{N_b} \sum_{i:\mathcal{B}(X_i)=b} Y_i \quad \text{and} \quad \widehat{V}_b := \frac{1}{N_b} \sum_{i:B(X_i)=b} (Y_i - \widehat{\pi}_b)^2$$

We now deploy an empirical Bernstein bound to produce a confidence interval for π_b.

Theorem 4. *For any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$,*

$$|\pi_b - \widehat{\pi}_b| \leq \sqrt{\frac{2\widehat{V}_b \log(3B/\alpha)}{N_b}} + \frac{3\log(3B/\alpha)}{N_b}, \quad \textit{simultaneously for all } b \in [B]$$

Corollary 3. *For $\alpha \in (0, 1)$, the function $h_n(\cdot) := \widehat{\pi}_{\mathcal{B}(\cdot)}$ is distribution-free (ε, α)-calibrated with*

$$\varepsilon = \sqrt{\frac{2\widehat{V}_b \log(3B/\alpha)}{N_{b^*}}} + \frac{3\log(3B/\alpha)}{N_{b^*}}$$

where $b^* = \arg\min_{b\in[B]} N_b$. Thus, $\{h_n\}_{n\in\mathbb{N}}$ is distribution-free asymptotically calibrated for any α.

This calibration guarantee can also be stated conditional on a given test point:

$$|\mathbb{E}[Y \mid f(X)] - f(X)| \leq \varepsilon \quad \text{almost surely } P_X$$

This holds since Theorem 4 provides simultaneously valid CIs for all regions $\mathcal{X}_b$.

4.2 Identifying a data-dependent partition using sample splitting

Binning uses a sample splitting strategy to learn the partition of $\mathcal{X}$ and estimate $\{\widehat{\pi}_b\}_{b\in[B]}$. The labeled data is split at random into a training data $\mathcal{D}_{\text{tr}}$ and a calibration set $\mathcal{D}_{\text{cal}}$. Then $\mathcal{D}_{\text{tr}}$ is used to train a scoring function $g : \mathcal{X} \to [0,1]$.

A binning scheme $\mathcal{B}$ is any partition of $[0,1]$ into B non-overlapping intervals $I_1, \cdots I_B$. $\mathcal{B}$ and g induce a partition of $\mathcal{X}$ as follows:

$$\mathcal{X}_b = \{x \in \mathcal{X} : g(x) \in I_b\},\ b \in [B]$$

The simplest binning scheme corresponds to fixed-width binning:

$$I_i = \left[\frac{i-1}{B}, \frac{i}{B}\right), i = 1, \cdots, B-1 \text{ and } I_B = \left[\frac{B-1}{B}, 1\right]$$

However, fixed-width binning suffers from low N_b. To remedy this, we consider uniform -mass binning. First, the calibration set $\mathcal{D}_{\text{cal}}$ is randomly split into two parts, $\mathcal{D}^1_{\text{cal}}$ and $\mathcal{D}^2_{\text{cal}}$. For $j \in [B-1]$, the (j/B)-th quantile of $g(X)$ is estimated from $\{g(X_i), i \in \mathcal{D}^1_{\text{cal}}\}$. Then, the bins are defined as:

$$I_1 = [0, \widehat{q}_1), I_i = [\widehat{q}_{i-1}, \widehat{q}_i], i = 2, \cdots, B-1 \text{ and } I_B = (\widehat{q}_{B-1}, 1]$$

where $\widehat{q}_j$ denotes the empirical quantile estimates. Now, only $\mathcal{D}^2_{\text{cal}}$ is used for calibrating the underlying classifier, as per thee calibration scheme defined in Section 4.1.

Theorem 5. *Fix $g : \mathcal{X} \to [0,1]$ and $\alpha \in (0,1)$. There exists a universal constant c such that if $|\mathcal{D}^1_{\text{cal}}| \geq cB\log(2B/\alpha)$, then with probability at least $1-\alpha$,*

$$N_{b^*} \geq |\mathcal{D}^2_{\text{cal}}|/2B - \sqrt{|\mathcal{D}^2_{\text{cal}}|\log(2B/\alpha)/2}$$

Thus even if $|\mathcal{D}^1_{\text{cal}}|$ does not grow with n, as long as $|\mathcal{D}^2_{\text{cal}}| = \Omega(n)$, uniform-mass binning is distribution-free $(\widetilde{O}(\sqrt{B\log(1/\alpha)/n}), \alpha)$-calibrated, and hence distribution-free asymptotically calibrated for any α.

If we use a small number of points (independent of n) for uniform-mass binning, and the rest to estimate bin probabilities, we achieve approximate/asymptotic distribution-free calibration.

4.3 Distribution-free calibration in the online setting

In online setting, the results of Section 4 are no longer valid since the number of calibration samples is unknown a priori and may even be dependent on the data. In order to quantify uncertainty in the online setting, we use time-uniform concentration bounds.

Fix a partition of $\mathcal{X}$, $\{\mathcal{X}_b\}_{b\in[B]}$. For some value of n, let the calibration data be given as $\mathcal{D}^{(n)}_{\text{cal}}$. Let $\{(X^b_i, Y^b_i)\}_{i\in[N^{(n)}_b]}$ be examples from the calibration set that fall into the partition $\mathcal{X}_b$, where $N^{(n)}_b := |\{i \in [n] : \mathcal{B}(X_i) = b\}|$. Let the empirical label average and cumulative (unnormalized) empirical variance be denoted as

$$\widehat{V}^+_b = 1 \vee \sum_{i=1}^{N^{(n)}_b} \left(Y^b_i - \overline{Y}^b_{i-1}\right)^2, \text{ where } \overline{Y}^b_i := \frac{1}{i}\sum_{j=1}^{i} Y^b_j \text{ for } i \in [N^{(n)}_b$$

Theorem 6. *For any $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$|\pi_b - \widehat{\pi}_b| \leq \frac{7\sqrt{\widehat{V}^+_b \log\left(1 + \log \widehat{V}^+_b\right)} + 5.3\log\left(\frac{6.3B}{\alpha}\right)}{N^{(n)}_b}, \quad \textit{simultaneously for all } b \in [B] \textit{ and all } n \in \mathbb{N}$$

In comparison to the batch setting, only a small price is paid for not knowing beforehand how many examples will be used for calibration.

4.4 Calibration under covariate shift

In covariate shift setting, calibration data $\{(X_i, Y_i)\}_{i\in[n]} \sim P^n$ is from a source distribution $P = P_X \times P_{Y|X}$, while the test point is from a shifted target distribution $(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}$. We assume the likelihood ratio (LR) $w : \mathcal{X} \to \mathbb{R}$; $w(x) := d\widetilde{P}_X(x)/dP_X(x)$ is well-defined. Even though w is arbitrarily ill-behaved and unknown, one can still make nontrivial claims using a modular approach towards assumptions:

Condition (A): $w(x)$ is known exactly and is bounded.

Condition (B): an asymptotically consistent estimator $\widehat{w}(x)$ for $w(x)$ can be constructed.

Under Condition (a), a weighted estimator using w delivers approximate and asymptotic distribution-free calibration. Under Condition (B), weighting with a plug-in estimator for w continues to deliver asymptotic distribution-free calibration. If Condition (B) is not met, then it implies that w is probably very ill-behaved, and so distribution-free calibration is also likely to be impossible.

For a fixed partition $\{\mathcal{X}_b\}_{b\in[B]}$, one can use the labeled data from the source distribution to estimate $\mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]$, given oracle access to w:

$$\check{\pi}_b^{(w)} := \frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}$$

Assume that for all $x \in \mathcal{X}$, $L \le w(x) \le U$ for some $0 < L \le 1 \le U < \infty$.

Theorem 7. *Assume w is known and bounded by L, U. Then for an explicit universal constant $c > 0$, with probability at least $1 - \alpha$,*

$$\left|\check{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]\right| \le c\left(\frac{U}{L}\right)^2 \sqrt{\frac{B\log(6B/\alpha)}{2n}}, \textit{simultaneously for all } b \in [B]$$

as long as $n \ge c(U/L)^2 B \log^2(6B/\alpha)$. Thus $h_n(\cdot) = \check{\pi}^{(w)}_{\mathcal{B}(\cdot)}$ is distribution-free asymptotically calibrated for any α.

Theorem 7 establishes distribution-free calibration under Condition (A). For Condition (B), using k unlabeled samples from the source and target domains, assume that we construct an estimator $\widehat{w}_k$ of w that is consistent. We now define an estimator

$$\check{\pi}_b^{(\widehat{w}_k)} := \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}$$

Proposition 2. *If $\widehat{w}_k$ is consistent, then $h_n(\cdot) = \check{\pi}^{(\widehat{w}_k)}_{\mathcal{B}(\cdot)}$ is distribution-free asymptotically calibrated for any $\alpha \in (0, 0.5)$.*

5 Other related work

The problem of assessing the calibration of binary classifiers was first studied in the meteorological and statistics literature. These works resulted in two common ways of measuring calibration: reliability diagrams and estimates of the squared expected calibration error (ECE). ECE is typically estimated through binning.

While these papers have focused on the difficulty of estimating calibration error, ours is the first formal impossibility result for achieving calibration. Our results show not only that binning is necessary for distribution-free calibration, but also sufficient.

6 Conclusion

We analyzed post-hoc uncertainty quantification for binary classification problems from the stand-point of robustness to distributional assumptions. By connecting calibration to confidence intervals and prediction sets, we established that popular parametric 'scaling' methods cannot provide informative calibration in the distribution-free setting. In contrast, we showed that a nonparametric 'binning' method satisfies approximate and asymptotic calibration guarantees without distributional assumptions. We also established guarantees for the cases of streaming dada and covariate shift.

7 Broader Impact

This paper explicitly relates the three aforementioned notions of uncertainty quantification without making distributional assumptions, describes what is possible and what is not. Importantly, by providing distribution-free guarantees on well-known variants of binning, we identify a conceptually simple and theoretically rigorous way to ensure calibration in high-risk real-world settings. Our tools are thus likely to lead to fairer systems, better estimates of risks of high-stakes decisions, and more human-interpretable outputs of classifiers that apply out-of-the-box in many real-world settings because of the assumption-free guarantees.

References

Rina Foygel Barber. Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2):3487-3524, 2020.

8 Question

1. Theorem 1 is more general? It holds in distribution-free setting? Theorem 2 is just for distribution-free setting.

2. The meaning of impossibility results are different? Barber's paper: CIs should be wide and it is not informative. In section 3.2, it is also similar. I think they are impossibility results for distribution-free CI and PS. But in section 3.3, it states the impossibility result for distribution-free asymptotic calibration. I was confused. Are they related?