Distribution-free possibilistic inference on conditional quantities

Leonardo Cella[1]

[1]Department of Statistical Sciences, Wake Forest University, USA

Abstract

Uncertainty quantification for conditional quantities—i.e., unknown quantities related to the conditional distribution of a response variable given covariates—is a fundamental problem. Existing methods often rely on restrictive parametric assumptions or smoothness conditions and typically only provide set estimates for the unknown quantities. This paper introduces Inferential Models (IMs) that offer possibilistic uncertainty quantification for conditional quantities, going beyond the simple provision of set estimates. Unlike traditional approaches, the proposed IMs are fully distribution-free and can handle both random and fixed conditional quantities. Moreover, they satisfy a marginal validity criterion, ensuring proper calibration of all IMs' outputs when averaged over the covariates distribution. Illustrations of this framework are provided for both random and fixed conditional quantities—specifically, a future response and the conditional median, respectively.

Keywords. inferential models, distribution-free, prediction, conditional median

1. Introduction

Consider a common scenario in which a study involves $p+1$ variables, $(X, Y) \subseteq \mathbb{R}^p \times \mathbb{R}$, where Y is the response variable—the primary focus—and the p covariates in X are used to *predict* or *understand* Y. Specifically, given observed pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ and a new covariate vector X_{n+1}, the goal is either to predict the next realization from the conditional distribution of Y given $X = x_{n+1}$ or to gain insight into this conditional distribution, such as by analyzing its moments or quantiles.

The assumption that $(X_1, Y_1), \dots, (X_n, Y_n)$ and X_{n+1} are independently drawn from the same distribution $\mathsf{P} = \mathsf{P}_{Y|X} \times \mathsf{P}_X$ is well-justified in most applications. However, it is often accompanied by a stronger, less defensible assumption that P follows a specific parametric model. In such cases, uncertainty quantification for an unknown quantity related to the conditional distribution of $Y \mid X = x$ becomes a two-step process: first, uncertainty is quantified for the model parameters, and then this uncertainty is propagated to the quantity of interest. This indirect approach can be problematic if the parametric model is misspecified. Therefore, in scenarios where there is little to no justification for assuming a parametric model, the ability to quantify uncertainty without imposing distributional assumptions is particularly valuable.

Quantifying uncertainty about the next response Y_{n+1} is often the primary goal in applications. A well-established approach for this is *conformal prediction* [2, 18, 27, 28], which constructs prediction sets with finite-sample coverage guarantees, all without relying on distributional assumptions on P. Yet, in many applications, extending uncertainty quantification beyond simple set estimates is desired. A more nuanced perspective involves assigning (potentially imprecise) probabilities to specific claims about Y_{n+1}.

Consider, for instance, a university evaluating an applicant with a high school GPA of 2.5 and other relevant characteristics. Rather than merely predicting a range for their future college GPA, the institution may be particularly concerned with probabilistically assessing a specific claim—say, whether the student's college GPA will fall below 2. Since such assessments influence important decisions, it is essential that probability assignments are properly calibrated, ensuring that erroneous conclusions remain controllably rare [26].

Inferential Models (IMs) [20, 23, 24] offer a possibilistic framework for uncertainty quantification about unknowns. Their key feature is the calibration of possibilistic assignments, ensuring that low possibility (or high necessity) measures are assigned to true (or false) claims at a controlled rate. Early IM developments focused on settings with an assumed parametric model, but recent advancements have extended their construction to distribution-free settings [4–10, 22]. In particular, the conditional prediction problem considered here was previously explored in Cella and Martin [9].

Challenges increase when the focus shifts from uncertainty quantification for the next realization of $\mathsf{P}_{Y|X}$ to uncertainty quantification for functionals of $\mathsf{P}_{Y|X}$. Traditional distribution-free methods aim to construct set estimates not just for a specific data point x_{n+1} but for all $x \in \mathbb{R}^p$. However, achieving proper coverage typically requires additional assumptions. Nonparametric regression [12, 29], for example, imposes smoothness conditions and assumes that $\mathsf{P}_{Y|X}$ has sufficiently light tails when estimating the conditional mean—reflecting the fundamental difficulty of constructing nontrivial confidence intervals for it [1]. Even robust functionals like conditional quantiles require additional continuity assumptions on the quantile function over $x \in \mathbb{R}^p$, as seen in methods such as quantile regression [17]. The same limitation applies to the distribution-free IMs construction for conditional quantiles proposed in Cella [5], which, while providing calibrated possibilistic uncertainty quantification, also relies on continuity assumptions.

This begs the question: can meaningful uncertainty quantification for at least some functionals of $\mathsf{P}_{Y|X}$ be achieved without imposing any additional assumptions? Recent advances in this area have been made by Medarametla and Candès [25] and Barber [3], both of which focus on uncertainty quantification through set estimates. A central element of these developments is the notion of validity they adopt. Similar to conformal prediction, the coverage of the proposed set estimates is considered marginally, rather than conditioned on the specific x_{n+1} observed. In other words, the goal is to cover the value of the unknown quantity of interest when weighted by the marginal distribution P_X. Since no assumptions are made about the class of distributions that P belongs to, it becomes more feasible to seek calibrated set estimates for individual data points, rather than attempting pointwise coverage across all $x \in \mathbb{R}^p$ [25].

Building on these advances, our goal is to extend these ideas within the IM framework, enabling distribution-free uncertainty quantification on conditional quantities beyond set estimates. In Section 4, we formalize the notion of marginal validity, discussed above, in the broader context of probabilistic uncertainty quantification and present a specific distribution-free IMs construction that satisfies this criterion. This construction is quite general, accommodating both random and fixed conditional quantities of interest. Specifically, we present two examples: one for a random quantity, Y_{n+1}, and another for a fixed quantity, the conditional median. The former was previously examined in Cella and Martin [9], but here, we reinterpret it through the lens of our new framework. The latter is a novel contribution, drawing significant inspiration from Medarametla and Candès [25].

The rest of the paper is organized as follows. In Section 2, a brief background on parametric IMs is provided to equip the reader with the foundational reasoning behind IMs, the logic of their construction, and the key properties they satisfy. Section 3 introduces (unconditional) distribution-free IMs, where the developments in Cella [5], originally designed for fixed quantities of interest, are generalized to accommodate both fixed and random quantities. This generalization is important because the IMs for conditional quantities proposed in Section 4 follow a similar logic. Finally, Section 5 provides a concise summary, key remarks, and a discussion of open problems.

2. Background on parametric IMs

Inferential Models (IMs) emerged a decade and a half ago with the intent of striking a balance between the two dominant schools of thought in statistics: the frequentist and Bayesian approaches. To put it another way, IMs were developed to provide probabilistic uncertainty quantification about unknowns, akin to the Bayesian framework, while ensuring that these probabilistic statements are calibrated and obtained without the need for prior specification, in line with the frequentist perspective. The key to achieving this "best of both worlds"—what Efron referred to as *the most unresolved problem in statistical inference* [14]—lies in the use of imprecise probabilities in IMs' probabilistic statements. In particular, these statements are expressed through necessity and possibility measures [13].

Consider a parametric model $\{\mathsf{P}_\theta : \theta \in \mathbb{T}\}$ consisting of probability distributions supported on a sample space $\mathbb{Z}$, indexed by a parameter space $\mathbb{T}$. Suppose that a random quantity Z, taking values in $\mathbb{Z}$, is of interest, and it is assumed that $Z \sim \mathsf{P}_\Theta$, where $\Theta \in \mathbb{T}$ is the unknown "true value" of the parameter. The observable data $Z^n = Z_1, \dots Z_n$ then consist of n i.i.d. realizations from P_Θ, so $Z^n \sim \mathsf{P}_\Theta^n$. The goal is to quantify uncertainty about Θ after observing data $Z^n = z^n$. Recent developments in Martin [21] construct an IM for Θ by applying a version of the *probability-to-possibility transform* [15, 16] to the model's relative likelihood

$$R(z^n, \theta) = \frac{L_{z^n}(\theta)}{\sup_\vartheta L_{z^n}(\vartheta)}.$$

More specifically, the IM's possibility contour is defined as

$$\pi_{z^n}(\theta) = \mathsf{P}_\theta^n\{R(Z^n, \theta) \le R(z^n, \theta)\}, \quad \theta \in \mathbb{T}, \tag{1}$$

and the possibility measure assigned to a claim C of the form $\Theta \in C$, where $C \subseteq \mathbb{T}$, is defined as

$$\overline{\Pi}_{z^n}(C) = \sup_{\theta \in C} \pi_{z^n}(\theta). \tag{2}$$

The corresponding necessity measure is defined via conjugacy: $\underline{\Pi}_{z^n}(C) = 1 - \overline{\Pi}_{z^n}(C^c)$.

As stated above, the frequentist calibration of these possibilistic assignments is the defining characteristic of IMs. In particular, the IMs' possibility measures are such that, for all $C \subseteq \mathbb{T}$,

$$\sup_{\Theta \in C} \mathsf{P}_\Theta^n\{\overline{\Pi}_{Z^n}(C) \le \alpha\} \le \alpha, \quad \text{for all } \alpha \in [0, 1].$$

In words, IMs assign small possibilities ($\leq \alpha$) to true claims at a small rate ($\leq \alpha$) as a function of data Z^n. A corresponding statement holds for the IMs' necessity measures, but we omit it here and throughout the paper. For a discussion of the important roles each measure plays in the IM framework, see Cella and Martin [11].

The foundation of these calibration properties is the so-called *validity property* of the IM contour, given by

$$\mathsf{P}^n_\Theta\{\pi_{Z^n}(\Theta) \leq \alpha\} \leq \alpha, \quad \text{for all } \alpha \in [0,1].$$

Several other key properties follow directly from this validity condition. To avoid redundancy, we defer their discussion to later sections

The reasoning behind the IM construction is as follows: the relative likelihood naturally quantifies the compatibility between a candidate value θ of the unknown parameter Θ and the observed data z^n. However, it does not inherently guarantee the calibrated probabilistic uncertainty quantification we seek. To address this, the relative likelihood undergoes the possibility-to-probability transformation in (1), a process termed *validification* by Martin [21]. This "compatibility function + validification" will be essential in developing the distribution-free IMs below.

3. Distribution-free IMs

The parametric IMs reviewed in Section 2 are powerful and arguably provide an appealing resolution to the long-standing frequentist vs. Bayesian debate in statistics. However, like all parametric methods, they come with a major drawback—the need to specify a parametric distribution for the data. In many modern applications, such assumptions often lack justification, making a distribution-free approach more desirable. Any serious framework for statistical inference must, therefore, be adaptable to distribution-free settings, and IMs are no exception.

Consider a random quantity of interest, Z, which takes values in a sample space $\mathbb{Z}$ and follows a distribution P, about which no assumptions are made. Observable data $Z^n = (Z_1, \dots, Z_n)$ consist of n i.i.d. realizations from P. The goal is to quantify uncertainty about an unknown quantity Θ that takes values in a space $\mathbb{T}$ and is associated with P, given the observed data $Z^n = z^n$. To keep things as general as possible, we consider cases where this unknown quantity is *fixed*, such as a functional of the underlying distribution $\Theta = \Theta(\mathsf{P})$ (e.g., a quantile of P), or *random*, such as a future realization $\Theta = Z_{n+1}$ from P. For the remainder of this section, $\mathbb{P}$ will represent probabilities taken over Z^n when Θ is fixed, and over both Z^n and Θ when Θ is random.

A key feature of the parametric IMs in Section 2 is their ability to make calibrated probabilistic assignments to any claims of interest. A distribution-free IM for Θ should uphold a similar property. More specifically, the goal is to construct a distribution-free IM that assigns small possibilities to true claims with small $\mathbb{P}$-probability. The following definition formalizes this requirement.

Definition 3.1. Let $\overline{\Pi}_{z^n}(C)$ denote the possibility measure assigned to a claim C of the form $\Theta \in C$, where $C \subseteq \mathbb{T}$. This possibility is said to be calibrated if it satisfies the following condition:

$$\mathbb{P}\big\{\overline{\Pi}_{Z^n}(C) \leq \alpha, \Theta \in C\big\} \leq \alpha \tag{3}$$

for all C, all $\alpha \in [0,1]$ and all distributions P on $Z \in \mathbb{Z}$.

For the construction of a distribution-free IM for Θ, the absence of an assumed model eliminates the likelihood function and, consequently, the relative likelihood. This makes it impossible to obtain a contour via the probability-to-possibility transform in (1). However, we argue that the core idea presented in Section 2 remains the same: validifying a real valued function that measures the compatibility between candidate values of Θ and the observed data z^n is sufficient for constructing distribution-free IMs. The key difference is that, in this setting, the compatibility function can no longer be based on the relative likelihood. A new strategy is needed.

Definition 3.2. Let $\rho : (\mathbb{Z}^n \times \mathbb{T}) \to \mathbb{R}$ be a function that quantifies the agreement between candidate values of Θ and the observed data z^n. This function is called a *compatibility-pivot* if it satisfies the following two conditions:

1. larger values of ρ indicate stronger agreement;
2. ρ is a pivot under $\mathbb{P}$, meaning it has a known distribution independent of any unknown quantities.

If, in a given application, such a compatibility-pivot can be identified, the possibility contour of the distribution-free IM for Θ is obtained via the probability-to-possibility transform of this function:

$$\begin{aligned} \pi_{z^n}(\theta) &= \mathbb{P}\{\rho(\mathcal{D}) \leq \rho(z^n, \theta)\} \\ &= F(\rho(z^n, \theta)), \quad \theta \in \mathbb{T}, \end{aligned} \tag{4}$$

where

$$\mathcal{D} = \begin{cases} (Z^n, \Theta) & \text{if } \Theta \text{ is random,} \\ (Z^n, \theta) & \text{if } \Theta \text{ is fixed,} \end{cases}$$

and F is the (known) cumulative distribution function (c.d.f.) of $\rho(\mathcal{D})$. The validity of this contour, along with the calibration of the corresponding possibility measure assigned to any claim C of the form $\Theta \in C$, as defined in (2), is established in the following theorem.

Theorem 3.1. *The contour in* (4) *satisfies the following validity property:*

$$\mathbb{P}\big\{\pi_{Z^n}(\Theta) \leq \alpha\big\} \leq \alpha, \tag{5}$$

for all $\alpha \in [0,1]$ and all distributions P on $Z \in \mathbb{Z}$. This implies that the corresponding possibility measure is calibrated in the sense of Definition 3.1.

Proof. The probability integral transform ensures that $F(\rho(Z^n, \Theta)) = \pi_{Z^n}(\Theta)$ is stochastically no smaller than Unif(0, 1), thereby establishing (5). For the second claim, as $\overline{\Pi}_{z^n}(C) = \sup_{\theta \in C} \pi_{z^n}(\theta)$, for any claim C such that $\Theta \in C$, $\overline{\Pi}_{Z^n}(C) \leq \alpha$ implies $\pi_{Z^n}(\Theta) \leq \alpha$, so (3) follows directly from (5). □

In addition to guaranteeing the calibration of the possibility measures assigned to claims of interest, (5) has two other important consequences. First, it implies that set estimates derived from the IM's contour have frequentist error rate control guarantees.

Corollary 3.1. *The α level sets of of the contour in (4) $C_\alpha(z^n) = \{\theta \in \mathbb{T} : \pi_{z^n}(\theta) > \alpha\}$ are nominal set estimates for Θ in the sense that*

$$\mathbb{P}\{C_\alpha(Z^n) \not\ni \Theta\} \leq \alpha,$$

for all $\alpha \in [0,1]$ and all distributions P on $Z \in \mathbb{Z}$.

Proof. It is clear that $C_\alpha(z^n) \not\ni \Theta$ if and only if $\pi_{Z^n}(\Theta) \leq \alpha$. From this, the result follows directly from (5). □

The second important consequence of (5) concerns the overall reliability of the IM's uncertainty quantification. In particular, it says that the calibration derived in Theorem 3.1 holds not just for some predetermined claims about Θ, but *uniformly* in said claims. Further discussion follows the theorem statement and proof.

Theorem 3.2. *The distribution-free IMs' possibility measure is uniformly calibrated in the sense that*

$$\mathbb{P}\{\overline{\Pi}_{Z^n}(C) \leq \alpha \text{ for some } C \text{ with } C \ni \Theta\} \leq \alpha, \quad (6)$$

for all $\alpha \in [0,1]$ and all distributions P on $Z \in \mathbb{Z}$.

Proof. Observe that a C satisfying $C \ni \Theta$ and $\overline{\Pi}_{z^n}(C) = \sup_{\theta \in C} \pi_{z^n}(\theta) \leq \alpha$ exists if and only if $\pi_{z^n}(\Theta) \leq \alpha$. The conclusion then follows immediately from (5), which guarantees that the probability of this event does not exceed α. □

The "for some C with $C \ni \Theta$" event in (6) can be seen as a union of every claim C that contains Θ. This is, of course, a much broader event than that associated with any single fixed C containing Θ, which implies that the probability bound in (6) is stronger than the analogous bound in (3). This stronger notion of calibration ensures that, even if the data analyst doesn't follow the recommended approach of setting a claim of interest prior to data collection and instead lets the data influence their choice of claim, erroneous conclusions remain controllably rare.

To illustrate the distribution-free IM construction presented above, we consider two examples. In both cases, $Z_1, \ldots, Z_n$ are i.i.d. continuous quantitative variables. The first example examines a fixed unknown quantity Θ, specifically the median of P. The second example focuses on a random unknown quantity Θ, namely the next realization Z_{n+1} of P.

Example 1. Consider the unknown quantity of interest to be the median of P, the exact point Θ such that $\mathsf{P}(Z_1 \leq \Theta) = 0.5$. As discussed in Cella [5], one natural choice for the compatibility-pivot in this context is

$$\rho(z^n, \theta) = \binom{n}{\gamma} 0.5^n, \quad (7)$$

where $\gamma = \sum_{i=1}^n I_{(0,\infty)}(z_i - \theta)$ represents the count of positive signs among $z_i - \theta$. This function, $\rho(z^n, \theta)$, corresponds to the probability mass function of a $\mathsf{Bin}(n, 0.5)$ random variable, so higher values indicate greater compatibility between z^n and θ. Moreover, it is also easy to see that $\rho(Z^n, \Theta)$ satisfies the second criterion for a compatibility-pivot in Definition 3.2, as it is simply a transformation of a $\mathsf{Bin}(n, 0.5)$ random variable.

As an illustration, consider a data set z^n with $n = 20$, whose histogram is displayed in the top panel of Figure 1. The dotted vertical line marks the sample median. In the bottom panel of Figure 1, the plausibility contour obtained from (7) is shown in black for this dataset, with its peak corresponding to the sample median.

Example 2. Consider the unknown quantity of interest to be a future realization $\Theta = Z_{n+1}$ of P. Define

$$T_i(Z^{n+1}) = T_i = -|Z_i - median(Z^{n+1}_{-i})|, \quad i = 1, \ldots, n+1,$$

where $Z^{n+1} = (Z^n, Z_{n+1}) = (Z^n, \Theta)$, and $Z^{n+1}_{-i} = Z^{n+1} \setminus \{Z_i\}$. Note that, while $T_1, \ldots, T_{n+1}$ are not i.i.d., they remain exchangeable due to the symmetry of the transformation $Z^{n+1} \to T^{n+1}$ and the exchangeability of $Z_1, \ldots, Z_{n+1}$, which follows from their i.i.d. nature.

An intuitive choice for the compatibility-pivot in this case is, therefore,

$$\rho(z^n, \theta) = r(T_{n+1}(z^{n+1})), \quad (8)$$

where $r(T_{n+1})$ denotes the rank of T_{n+1} relative to $T_1, \ldots, T_n$. Notably, (8) increases when the candidate value θ for the next $\Theta = Z_{n+1}$ aligns well with the observed data, resulting in a higher rank. Furthermore, the exchangeability of the T_is and the fact that they are continuous ensure that $r(T_{n+1})$ is uniformly distributed over $\{1, 2, \ldots, n+1\}$, i.e.,

$$\rho(Z^n, \Theta) \sim \mathsf{Unif}\{1, 2, \ldots, n+1\},$$

confirming that it serves as a compatibility-pivot. From (4), the distribution-free IM contour for $\Theta = Z_{n+1}$ is given by

$$\pi_{z^n}(\theta) = \mathbb{P}\{\mathsf{Unif}\{1, \ldots, n+1\} \leq r(T_{n+1}(z^{n+1}))\}$$

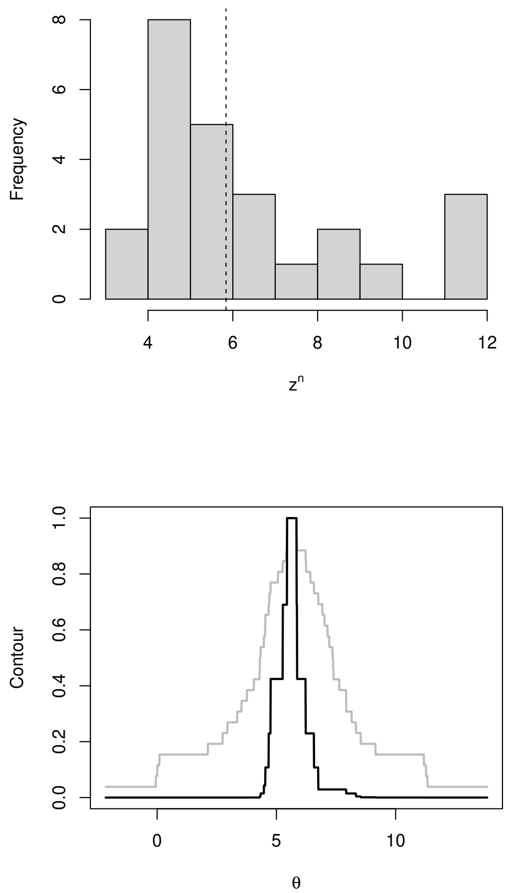


Figure 1. *Illustrations for the examples in Section 3.*

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} 1\{T_i(z^{n+1}) \le T_{n+1}(z^{n+1})\}.$$

The bottom panel of Figure 1 displays, in grey, this contour for the data shown in the histogram of the top panel. Since T_{n+1} is based on the median of z^n, the contour peaks at the sample median. It's also worth noting the greater precision of the contour for the median of P, as derived in Example 1. This makes sense, as quantifying uncertainty for a random quantity like Z_{n+1} is inherently more complex than for a fixed quantity like the median; but see Section 4.4.

4. Distribution-free IMs for conditional quantities

4.1. Setup and goals. In this section, which addresses the main problem of the present paper, data $Z^n = (Z_1, \dots, Z_n)$ consist of n i.i.d realizations of $Z = (X, Y) \subseteq \mathbb{R}^p \times \mathbb{R}$, where X denotes the covariates and Y represents a quantitative response. No assumptions are made about the distribution $\mathsf{P} = \mathsf{P}_{Y|X} \times \mathsf{P}_X$ of Z. The focus is on an unknown quantity Θ related to $\mathsf{P}_{Y|X}$. More specifically, after observing data Z^n and a new vector of covariates X_{n+1}, the unknown quantity of interest is related to the distribution of $Y \mid X_{n+1} = x_{n+1}$. Similar to Section 3, Θ can be random, such as the next realization Y_{n+1} of the conditional distribution, or fixed, such as its median.

To construct an IM for these conditional distribution-related Θ's, we first need to clearly define the type of validity we aim to achieve. The observed data consists of the n covariate-response pairs in Z^n, along with the $(n+1)$-th covariate vector X_{n+1}. We denote this combined data by W_{n+1}, i.e., $W_{n+1} = \{Z^n, X_{n+1}\}$. The following definition specifies the desired validity that the contour of a distribution-free IM for Θ based on $W_{n+1} = w_{n+1}$ should satisfy, while the subsequent theorem outlines the properties of an IM with such a contour. Throughout this section, $\mathbb{P}$ denotes the probability taken over W_{n+1} when Θ is fixed, and over both W_{n+1} and Θ when Θ is random.

Definition 4.1. A distribution-free IM for Θ with contour $\pi_{w^{n+1}}$ is valid if

$$\mathbb{P}\{\pi_{W_{n+1}}(\Theta) \le \alpha\} \le \alpha, \tag{9}$$

for all $\alpha \in [0,1]$ and all distributions P on $Z \in \mathbb{R}^p \times \mathbb{R}$.

Theorem 4.1. *For a distribution-free IM's contour that satisfies (9), the following is true for all $\alpha \in [0,1]$ and all distributions P on $Z \in \mathbb{R}^p \times \mathbb{R}$:*

1. *The possibility measures for all claims $C \subseteq \mathbb{T}$ are such that*
$$\mathbb{P}\{\overline{\Pi}_{W^{n+1}}(C) \le \alpha, \Theta \in C\} \le \alpha. \tag{10}$$
2. *The α level sets of of the IM's contour*
$$C_\alpha(w^{n+1}) = \{\theta \in \mathbb{T} : \pi_{w^{n+1}}(\theta) > \alpha\} \tag{11}$$
are nominal set estimates for Θ in the sense that
$$\mathbb{P}\{C_\alpha(W^{n+1}) \not\ni \Theta\} \le \alpha.$$
3. *The possibility measures are uniformly calibrated in the sense that*
$$\mathbb{P}\{\overline{\Pi}_{W^{n+1}}(C) \le \alpha \text{ for some } C \text{ with } C \ni \Theta\} \le \alpha.$$

Proof. The following hold for the corresponding statements above:

1. For any claim C that contains Θ we have that $\overline{\Pi}_{w^{n+1}}(C) \ge \pi_{w^{n+1}}(\Theta)$ due to the monotonicity of the possibility contour.
2. $C_\alpha(w^{n+1}) \not\ni \Theta$ if and only if $\pi_{w^{n+1}}(\Theta) \le \alpha$.
3. There exists a C such that $C \ni \Theta$ and $\overline{\Pi}_{w^{n+1}}(C) = \sup_{\theta \in C} \pi_{w^{n+1}}(\Theta) \le \alpha$ if and only if $\pi_{z^n}(\Theta) \le \alpha$.

These facts, combined with (9), give the results. $\square$

Although our primary interest lies in conditional quantities, the validity under consideration here is marginal, meaning it is not conditioned on the observed x_{n+1} but rather averaged over all possible values of X_{n+1}. This is most easily understood in the context of set estimates. The conditional coverage probability of a set estimate is given by

$$x_{n+1} \mapsto \mathbb{P}\{C_\alpha(Z^n, x_{n+1}) \ni \Theta \mid X_{n+1} = x_{n+1}\},$$

which is a function of x_{n+1}. The validity property considered here ensures that the expected value of this function, taken with respect to P_X, is at least $1-\alpha$. This, however, does not provide any information about the conditional coverage for any particular value of x_{n+1}. Similarly, conditional calibration of the IM's possibility measure would mean

$$\mathbb{P}\{\overline{\Pi}_{(Z^n, x_{n+1})}(C) \le \alpha, \Theta \in C \mid X_{n+1} = x_{n+1}\} \le \alpha,$$

for all $x_{n+1} \in \mathbb{R}^p$, but the notion in (10) averages the left hand side above across P_X. Achieving conditional validity in a distribution-free setting, where no assumptions are made about the class of distributions to which P belongs, is highly challenging yet practically significant. We briefly discuss this in Section 5.

But how can we construct a distribution-free IM whose contour satisfies (9)? The approach we propose closely follows the construction outlined in Section 3.

4.2. IMs construction. The IMs construction begins with the specification of a compatibility-pivot, in the same spirit as in Definition 3.2, but adapted to the present context. More specifically, let $\rho : (\mathbb{W}^{n+1} \times \mathbb{T}) \to \mathbb{R}$ be a function that measures the level of agreement between candidate values of the unknown quantity of interest, Θ, and the observed data w^{n+1}. Assume that larger values of this function correspond to stronger agreement. Furthermore, suppose ρ is as a pivot, meaning its distribution, when considered as a function of (W^{n+1}, Θ), is known and independent of any unknown quantities.

Theorem 4.2. *The contour*

$$\begin{aligned} \pi_{w^{n+1}}(\theta) &= \mathbb{P}\{\rho(\mathcal{D}) \le \rho(w^{n+1}, \theta)\} \\ &= F(\rho(w^{n+1}, \theta)), \quad \theta \in \mathbb{T}, \end{aligned} \tag{12}$$

where

$$\mathcal{D} = \begin{cases} (W^{n+1}, \Theta) & \text{if } \Theta \text{ is random,} \\ (W^{n+1}, \theta) & \text{if } \Theta \text{ is fixed,} \end{cases}$$

and F is the c.d.f of $\rho(\mathcal{D})$, is valid in the sense of Definition 4.1.

Proof. The proof is identical to the one in Theorem 3.1. The result follows from the fact that $F(\rho(W^{n+1}, \Theta))$ is stochastically no smaller than $\mathsf{Unif}(0, 1)$. $\square$

The distribution-free IM for Θ, with a contour defined as in (12), retains all the properties established in Theorem 4.1. A key challenge, however, is identifying a suitable compatibility-pivot ρ, which plays a crucial role in the proposed construction. This challenge is similar to that in the (unconditional) distribution-free IM construction in Section 3. While specific applications allow for its identification, as illustrated in the examples below, a broadly applicable strategy remains elusive. In fact, such compatibility-pivots may not always exist—see Section 5 for further discussion.

4.3. IMs for conditional prediction. Consider the case where, after observing data $W^{n+1} = (Z^n, X_{n+1})$, the goal is to quantify uncertainty about the next response, i.e., $\Theta = Y_{n+1}$. The distribution-free IM construction above requires identifying a compatibility-pivot function $\rho(w^{n+1}, \theta)$ that not only measures the agreement between a candidate value θ for the next response Θ and the observed data w^{n+1}, but also has a known distribution, independent of unknowns, when viewed as a function of W^{n+1} and Θ.

In the search of such compatibility-pivot, recall the IMs solution to the (unconditional) prediction problem discussed in Section 3, Example 2. To mirror that approach here, define

$$T_i(Z^{n+1}) = T_i = -\left|Y_i - \hat{m}_{-i}^{n+1}(X_i)\right|, \quad i = 1, \dots, n+1,$$

where $Z^{n+1} = (Z^n, (X_{n+1}, Y_{n+1})) = (Z^n, (X_{n+1}, \Theta))$ and $\hat{m}_{-i}^{n+1}$ can be any regression algorithm, e.g., a median regression model, fitted with data $Z^{n+1} \setminus \{X_i, Y_i\}$.

The same argument from Example 2 in Section 3 regarding the exchangeability of $T_1, \dots, T_{n+1}$ applies here. Specifically, the exchangeability of Z^{n+1} is preserved due to the symmetry of the transformation $Z^{n+1} \to T^{n+1}$. This implies that the rank of T_{n+1}, denoted $r(T_{n+1})$, relative to $T_1, \dots, T_n$, is equally likely to take any value between 1 and $n+1$. We can then define the compatibility-pivot as

$$\rho(w^{n+1}, \theta) = r(T_{n+1}(z^{n+1})),$$

and the fact that $\rho(W^{n+1}, \Theta) \sim \mathrm{Unif}\{1, \dots, n+1\}$ leads to the distribution-free IM contour for Θ given by

$$\begin{aligned} \pi_{w^{n+1}}(\theta) &= \mathbb{P}\left\{\mathrm{Unif}\{1, \dots, n+1\} \le r(T_{n+1}(z^{n+1}))\right\} \\ &= \frac{1}{n+1} \sum_{i=1}^{n+1} 1\{T_i(z^{n+1}) \le T_{n+1}(z^{n+1})\}. \end{aligned} \tag{13}$$

As an illustration, consider the dataset z^n with $n = 200$ in the first plot of Figure 2. The line in the graph represents the third-degree median regression model selected to fit the data z^n. Interest, however, is in Y_{n+1}, but the third degree median regression model like the one displayed is what we will use as $\hat{m}$ in the construction above. Suppose that $X_{n+1} = 7$ is observed. The bottom graph in Figure 2 displays the contour in (13). The

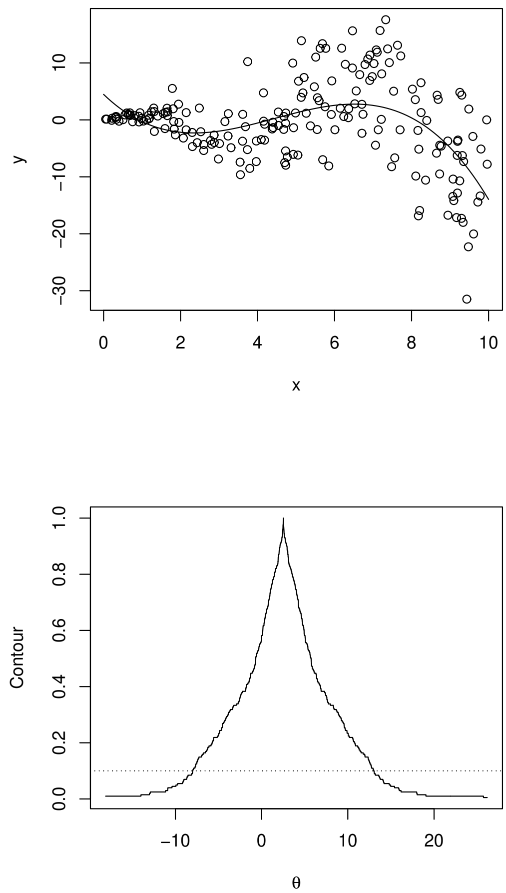


Figure 2. *Illustrations for the example in Section 4.3.*

horizontal line determines the corresponding 95% set estimate for Y_{n+1} derived by (11).

4.4. IMs for the conditional median. We now shift focus to a fixed Θ, specifically the conditional median. Denote by $m(x)$ the median of the conditional distribution of Y given $X = x$. Given the observed data w^{n+1}, the goal is to construct a distribution-free IM for $\Theta = m(x_{n+1})$.

To achieve this, a suitable compatibility-pivot must be identified. While one might be tempted to draw inspiration from the (unconditional) median example in Section 3, that approach won't work here for the simple reason that when X includes at least one quantitative variable, there will be no replicates for Y given $X = x$. In Cella [5], this issue was addressed by creating neighborhoods of X, which enabled the construction of a distribution-free IM for the conditional median based on the additional continuity assumption of the median function over $x \in \mathbb{R}^p$. Since we are not willing to assume continuity in the present paper, a new strategy must be developed.

One could also try to draw inspiration from the conditional prediction problem discussed in Section 4.3 above, where the unknown quantity of interest is Y_{n+1}. In that case, the exchangeability of (X_{n+1}, Y_{n+1}) with $Z_1, \dots, Z_n$ was crucial in defining a compatibility-pivot. One might consider using a similar reasoning if $(X_{n+1}, m(X_{n+1}))$ were exchangeable with $Z_1, \dots, Z_n$, but clearly, this is not the case. However, exchangeability-related ideas will be crucial to our proposed solution.

We begin by randomly splitting the observable data Z^n into two subsets, Z^{n_1} and Z^{n_2}, with sizes n_1 and n_2, respectively, such that $n_1 + n_2 = n$. Note that the observable X_{n+1} is not included in this split. Next, a median regression model $\hat{m}$ is fitted using Z^{n_1}. There is no restriction on the type of model here; the selection is entirely at the discretion of the data analyst.

Now, let (X_i, Y_i) represent the covariates/response for the i-th observation in Z^{n_2}, and relabel X_{n+1}, the observed covariates that were not included in the data split above, as X_{n_2+1}. Define

$$M_i = -\left|m(X_i) - \hat{m}(X_i)\right|, \quad i = 1, \dots, n_2 + 1.$$

Note that

$$\begin{aligned} M_{n_2+1} &= -\left|m(X_{n_2+1}) - \hat{m}(X_{n_2+1})\right| \\ &= -\left|\Theta - \hat{m}(X_{n+1})\right| \end{aligned}$$

is the M which Θ belongs to. The following lemma establishes the distribution of the rank of M_{n_2+1} with respect to $M_1, \dots, M_{n_2}$.

Lemma 4.1. *The rank of M_{n_2+1} with respect to $M_1, \dots, M_{n_2}$, which we will denote by $r(M_{n_2+1})$, is such that*

$$r(M_{n_2+1}) \sim \mathsf{Unif}\{1, \dots, n_2 + 1\}.$$

Proof. Since Z^n are i.i.d., X_{n_2+1} is independent of Z^n, and $\hat{m}$ is fitted using only Z^{n_1}, it follows that $M_1, \dots, M_{n_2+1}$ are i.i.d. as well, which in turn implies their exchangeability. Consequently, their ordering is uniformly random, establishing the result. □

If we had access to $m(X_i)$ for $i = 1, \dots, n_2$, we could directly apply the previous result to define our compatibility-pivot and complete the distribution-free IM construction for the conditional median. This approach is infeasible in practice, however, because $m(X_i)$ for $i = 1, \dots, n_2$ are unobservable. The only quantities available to us for each X_i in Z^{n_2} are the corresponding responses Y_i. Define then

$$T_i = -\left|Y_i - \hat{m}(X_i)\right|, \quad i = 1, \dots, n_2, \tag{14}$$

and let $r^*(M_{n_2+1})$ to be the rank of M_{n_2+1} with respect to $T_1, \dots, T_{n_2}$. Since $T_1, \dots, T_{n_2}$ are computable, determining the distribution of $r^*(M_{n_2+1})$ and verifying its independence from any unknown quantities allows us to use this rank as the compatibility-pivot. The following lemma is crucial for deriving the distribution of $r^*(M_{n_2+1})$, which is formally presented in the subsequent theorem.

Lemma 4.2. *For $i = 1, \dots, n_2$,*

$$\mathbb{P}\{T_i \le M_i\} \ge 1/2.$$

Proof. By definition of the conditional median, we have $\mathbb{P}\{Y_i \ge m(X_i)\} \ge 1/2$ and $\mathbb{P}\{Y_i \le m(X_i)\} \ge 1/2$. Fix X_i and consider the following two cases:

1. $m(X_i) \ge \hat{m}(X_i)$;
2. $m(X_i) < \hat{m}(X_i)$.

In the first case, since the events $\{Y_i \ge m(X_i)\}$ and $\{m(X_i) \ge \hat{m}(X_i)\}$ are independent, it follows that the probability of

$$M_i = \hat{m}(X_i) - m(X_i) \ge T_i = \hat{m}(X_i) - Y_i$$

is at least $1/2$. Similarly, in the second case, the independence of the events $\{Y_i \le m(X_i)\}$ and $\{m(X_i) \le \hat{m}(X_i)\}$ implies that the probability of

$$M_i = m(X_i) - \hat{m}(X_i) \ge T_i = Y_i - \hat{m}(X_i)$$

is also at least $1/2$. Since the conclusion holds conditionally in both cases, marginalizing out X_i establishes the result. □

Theorem 4.3. *For $t = 1, \dots, n_2 + 1$,*

$$\mathbb{P}\{r^*(M_{n_2+1}) \le t\} \le 1 - \sum_{j=1}^{n_2+1} \frac{1}{n_2+1} \sum_{k=t}^{j-1} \binom{j-1}{k} 0.5^{j-1}. \tag{15}$$

Proof. Let $N(T) = \#\{i = 1, \dots, n_2 : T_i \le M_{n_2+1}\}$. For $m = 0, \dots, n_2$, by the law of total probabilities,

$$\mathbb{P}\{N(T) \ge m\} = \sum_{j=1}^{n_2+1} \mathbb{P}\{N(T) \ge m \mid r(M_{n_2+1}) = j\} \times \mathbb{P}\{r(M_{n_2+1}) = j\}.$$

Lemma 4.1 implies that $\mathbb{P}\{r(M_{n_2+1}) = j\} = \frac{1}{n_2+1}$ for all $j = 1, \dots, n_2 + 1$. Now, for a given j, note that if $r(M_{n_2+1}) = j$, then $j - 1$ out of the n_2 values in $M_1, \dots, M_{n_2}$ are less than M_{n_2+1}. This, together with Lemma 4.2, implies that $\mathbb{P}\{N(T) \ge m \mid r(M_{n_2+1}) = j\}$ can be bounded below by the probability of at least m successes of a $\mathsf{Bin}(j-1, 0.5)$ random variable, so

$$\mathbb{P}\{N(T) \ge m\} \ge \sum_{j=1}^{n_2+1} \frac{1}{n_2+1} \sum_{k=m}^{j-1} \binom{j-1}{k} 0.5^{j-1}.$$

The fact that $\mathbb{P}\{r^*(M_{n_2+1}) \le t\} = \mathbb{P}\{N(T) \le t-1\}$ establishes the result. □

Despite the fact that we could not derive the exact distribution of $r^*(M_{n_2+1})$, the upper bound in (15) is sufficient to compute a valid distribution-free IM's contour for the conditional median. More specifically,

$$\pi_{w^{n+1}}(\theta) = 1 - \sum_{j=1}^{n_2+1} \frac{1}{n_2+1} \sum_{k=r^*_{obs}}^{j-1} \binom{j-1}{k} 0.5^{j-1}, \tag{16}$$

where $r^*_{obs} = r^*_{obs}(M_{n_2+1})$ is $r^*(M_{n_2+1})$ calculated from the observed data w^{n+1}, is the possibility contour of a valid distribution-free IM for $\Theta = m(x_{n_2+1})$.

Theorem 4.4. *The contour in (16) for $\Theta = m(x_{n_2+1})$ is valid in the sense of Definition 4.1.*

Proof. Define $G(t) = \mathbb{P}\{r^*(M_{n_2+1}) \le t\}$ and let

$$\dot{\pi}_{w^{n+1}}(\theta) = G\left(r^*_{obs}(M_{n_2+1})\right),$$

which represents the IM contour based on the exact distribution of $r^*(M_{n_2+1})$. Theorem 4.1 guarantees the validity of this contour, ensuring that it satisfies (9). Since $\pi_{w^{n+1}}(\theta)$ in (16) is always at least $\dot{\pi}_{w^{n+1}}(\theta)$ for all $\theta \in \mathbb{T}$, it follows directly that $\pi_{w^{n+1}}(\theta)$ also satisfies (9). □

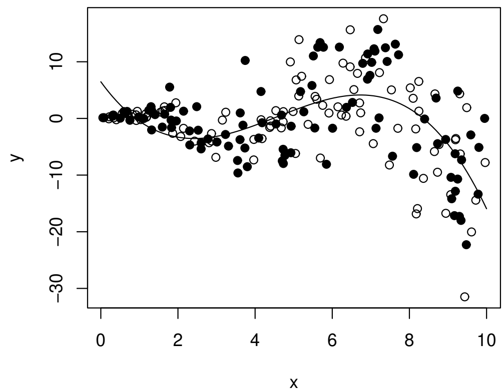
Consider again the dataset z^n with $n = 200$ in the top panel of Figure 2. The heteroskedasticity is evident, as the dispersion of Y varies considerably with X. This suggests that the median is an appropriate measure of central tendency for the conditional distribution of Y given X. Now, suppose $X_{n+1} = 7$. To apply the distribution-free IM construction for $\Theta = m(x_{n+1})$ described above, we first randomly split the data z^n into two halves, with z^{n_1} represented by the black data points in the top panel of Figure 3. The line in the same graph represents the third-degree median regression model chosen for $\hat{m}$. The bottom panel of Figure 3 displays the contour from (16) in black. For comparison, the contour for $\Theta = Y_{n+1}$ is displayed in grey. It's important to note that this is a different contour than the one displayed in the bottom panel of Figure 2. To ensure a fair comparison, the contour here was derived using a modified version of the construction presented in Section 4.3. This approach relies on data-splitting and leverages the exchangeability of $T_1, \dots, T_{n_2}$ from (14) and

$$T_{n_2+1} = -|Y_{n_2+1} - \hat{m}(X_{n_2+1})|,$$

with Y_{n+1} being relabeled Y_{n_2+1}. It leads to the following contour:

$$\pi_{w^{n+1}}(\theta) = \frac{1}{n_2+1} \sum_{i=1}^{n_2+1} 1\{T_i(z^{n+1}) \le T_{n_2+1}(z^{n+1})\}.$$

This contour proves to be significantly more precise than the one for the conditional median, a stark contrast to the



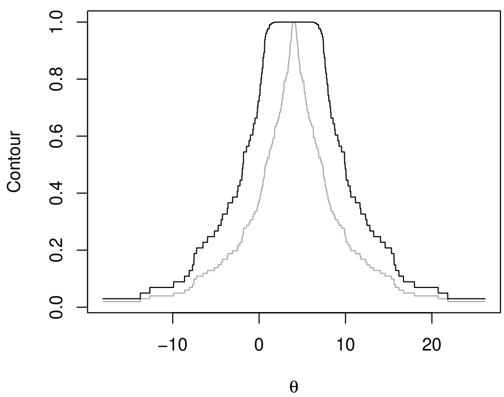


Figure 3. *Illustrations for the example in Section 4.4.*

examples in Section 3, where the opposite trend was observed. Moreover, comparing the right-hand side above to that in (16) reveals that the former is indeed smaller than the latter. Further discussion on this appears in Section 5.

5. Conclusion

This paper introduces a novel distribution-free IMs construction tailored for possibilistic inference on conditional quantities—those linked to the conditional distribution of a response given the covariates. The proposed method is versatile, capable of handling both random and fixed unknown quantities, and it is provably valid, as demonstrated in Theorem 4.1. The concept of validity used here is marginal, which contrasts with, and is weaker than, the more commonly discussed conditional validity property. Achieving pointwise validity across the entire conditional distribution in a distribution-free context is exceedingly difficult without additional assumptions. Moreover, based on results such as those in Lei and Wasserman [19], it seems improbable that any non-trivial IM could satisfy conditional validity.

This section concludes with some remarks and suggestions for future work. First, although our proposed construction is broadly applicable, its implementation depends heavily on identifying a suitable compatibility-pivot. No universal strategy exists for this task, meaning each case must be approached individually. Additionally, it's not always guaranteed that a suitable pivot will be found, particularly for quantities influenced by the tails of the distribution. One natural direction for future research is to apply our method to other conditional quantities where it may be possible to identify compatibility-pivots. For example, the approach used for the conditional median could be extended to conditional quantiles. Other quantities include the conditional interquartile range, trimmed mean, and proportion. It is also important to mention that using compatibility pivots to construct distribution-free IMs is not the only approach. Other strategies exist and can be employed with minimal or no loss of validity; see [10, 22].

Second, the proposed solution for the conditional median relied on data splitting, which was crucial for approximating the distribution of the selected compatibility-pivot. However, similar to how we can choose whether or not to use data-splitting when constructing an IM for Y_{n+1}, alternative versions of the approach presented here may eliminate the need for data-splitting in the conditional median case. This offers a promising direction for future research.

Lastly, we observed in the example of Section 4.4 that the proposed IM for the conditional median was less efficient than the IM for Y_{n+1}, which might initially seem counterintuitive and could suggest that our approach is suboptimal. On the one hand, it's important to recall that our solution relied on an upper bound of the true distribution function of the chosen compatibility-pivot, so there may indeed be room for improvement. On the other hand, Medarametla and Candès [25] demonstrates that confidence intervals containing Y_{n+1} with probability $1 - \alpha/2$ must also contain the conditional median with probability $1 - \alpha$, suggesting that what we observed in our example is not entirely unexpected. Further exploration of these findings within the context of imprecise probabilities will be conducted and reported in a future extension of this manuscript.

Additional Author Information

Acknowledgements. The author thanks the three anonymous reviewers for their valuable feedback on an earlier version of this manuscript.

References

[1] R. R. Bahadur and L. J. Savage. "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems". In: *The Annals of Mathematical Statistics* 27.4 (1956), pp. 1115–1122.

[2] V. Balasubramanian, S.S. Ho, and V. Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Elsevier Science, 2014.

[3] R. F. Barber. "Is distribution-free inference possible for binary regression?" In: *Electronic Journal of Statistics* 14.2 (2020), pp. 3487–3524.

[4] L. Cella. "Finite sample valid probabilistic inference on quantile regression". In: *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by Enrique Miranda, Ignacio Montes, Erik Quaeghebeur, and Barbara Vantaggi. Vol. 215. Proceedings of Machine Learning Research. PMLR, Nov. 2023, pp. 109–118.

[5] L. Cella. "Distribution-free Inferential Models: Achieving finite-sample valid probabilistic inference, with emphasis on quantile regression". In: *International Journal of Approximate Reasoning* 170 (2024), p. 109211.

[6] L. Cella and R. Martin. "Approximately Valid and Model-Free Possibilistic Inference". In: *Belief Functions: Theory and Applications*. Ed. by Thierry Denœux, Eric Lefèvre, Zhunga Liu, and Frédéric Pichon. Cham: Springer International Publishing, 2021, pp. 127–136.

[7] L. Cella and R. Martin. "Valid inferential models for prediction in supervised learning problems". In: *Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by Andrés Cano, Jasper De Bock, Enrique Miranda, and Serafín Moral. Vol. 147. Proceedings of Machine Learning Research. PMLR, June 2021, pp. 72–82.

[8] L. Cella and R. Martin. "Validity, consonant plausibility measures, and conformal prediction". In: *International Journal of Approximate Reasoning* 141 (2022), pp. 110–130.

[9] L. Cella and R. Martin. "Valid inferential models for prediction in supervised learning problems". In: *International Journal of Approximate Reasoning* 150 (2022), pp. 1–18.

[10] L. Cella and R. Martin. "Direct and approximately valid probabilistic inference on a class of statistical functionals". In: *International Journal of Approximate Reasoning* 151 (2022), pp. 205–224.

[11] L. Cella and R. Martin. "Possibility-theoretic statistical inference offers performance and probativeness assurances". In: *International Journal of Approximate Reasoning* 163 (2023), p. 109060.

[12] W.J. Conover. *Practical Nonparametric Statistics*. Wiley, 1971.

[13] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, 1988, pp. xvi+263.

[14] B. Efron. "Discussion: "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review"". In: *International Statistical Review* 81.1 (2013), pp. 41–42.

[15] D. Hose and M. Hanss. "On data-based estimation of possibility distributions". In: *Fuzzy Sets and Systems* 399 (2020). Fuzzy Intervals, pp. 77–94.

[16] D. Hose and M. Hanss. "A universal approach to imprecise probabilities in possibility theory". In: *International Journal of Approximate Reasoning* 133 (2021), pp. 133–158.

[17] R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005, pp. xvi+349.

[18] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. "Distribution-Free Predictive Inference for Regression". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111.

[19] J. Lei and L. Wasserman. "Distribution-free Prediction Bands for Non-parametric Regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76.1 (July 2013), pp. 71–96.

[20] R. Martin. "False confidence, non-additive beliefs, and valid statistical inference". In: *International Journal of Approximate Reasoning* 113 (2019), pp. 39–73.

[21] R. Martin. "Valid and efficient imprecise-probabilistic inference with partial priors, I. First results." 2022. arXiv: 2203.06703.

[22] R. Martin. "Valid and efficient imprecise-probabilistic inference with partial priors, III. Marginalization". 2023. arXiv: 2309.13454.

[23] R. Martin and C. Liu. "Inferential Models: A framework for prior-free posterior probabilistic inference". In: *Journal of the American Statistical Association* 108 (2013), pp. 301–313.

[24] R. Martin and C. Liu. *Inferential Models: Reasoning with Uncertainty*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press, 2015.

[25] D. Medarametla and E. Candès. "Distribution-free conditional median inference". In: *Electronic Journal of Statistics* 15.2 (2021), pp. 4625–4658.

[26] N. Reid and D. R. Cox. "On some principles of statistical inference". In: *International Statistical Review* 83.2 (2015), pp. 293–308.

[27] G. Shafer and V. Vovk. "A Tutorial on Conformal Prediction". In: *Journal of Machine Learning Research* 9 (2007), pp. 371–421.

[28] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Berlin, Heidelberg: Springer-Verlag, 2005.

[29] L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York, 2006.