# A Simplified Derivation of Confidence Regions Based on Inferential Models

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
vladik@utep.edu

**Abstract**

Recently, a new *inferential models* approach has been proposed for statistics. Specifically, this approach provides a new random-set-based way to come up with confidence regions. In this paper, we show that the confidence regions obtained by using the main version of this new methodology can also be naturally obtained directly, without invoking random sets.

## 1 Finding Confidence Regions: Formulation of the Problem

**Finding the confidence interval.** We have a family of distributions $f_\theta(x)$ characterized by a parameter (or parameters) $\theta \in \Theta$. We have a sample $x_1, \ldots, x_n$ from a distribution $f_\theta(x)$ corresponding to some unknown value of this parameter. Our task is to extract information about $\theta$ from the sample.

Let $\alpha > 0$ be a real number. A function $C$ that maps a sample $x = (x_1, \ldots, x_n)$ to subsets of the set $\Theta$ is called a *confidence region* $C(x)$ if for every $\theta \in \Theta$, the actual value $\theta$ is contained in the set $C(x)$ with probability $\geq 1 - \alpha$ (see, e.g., [5]):

$$\mathrm{Prob}(\theta \in C(x)) \geq 1 - \alpha.$$

*Comment.* Often, confidence regions are formed based on a sufficient statistic $s(x)$.

**Inferential models approach to finding confidence regions.** Recently, a new *inferential models* approach has been proposed for designing confidence regions; see, e.g., [1, 2, 3] and references therein. This approach based on random sets; see, e.g., [4].

**What we do in this paper.** In this paper, we show that the confidence regions obtained by using the main version of the inferential models approach can also be derived in a straightforward way, without a need to invoke random sets.

## 2 How Confidence Regions Are Designed in Inferential Models Approach: A Brief Reminder

**First step of the inferential models approach: general idea.** The inferential model approach start with representing the available statistical information – in particular, information about the value $s$ of the sufficient statistic $s(x)$ – as $s = a(\theta, U)$, where $U$ is a random variable with a known probability distribution.

This formula for the random variable $s$ is called an *association*.

**First step of the inferential models approach: main version.** In the main version, as the variable $U$, the authors of the inferential models approach propose to take a variable uniformly distributed on the interval $[0, 1]$.

For such $U$, to represent a general probability distribution in the desired form, we can use the known fact that each such variable, with the cumulative distribution function (cdf) $F(z)$, can be represented as $F^{-1}(U)$, where $F^{-1}(z)$ denotes an inverse function. This easy-to-check fact is one of the main ways to simulate random variables.

In our case, the distribution of $s$ (depending on $\theta$) has the cdf $G_\theta(s)$. Thus, the corresponding model has the form $s = G_\theta^{-1}(U)$.

**Second step of the inferential models approach: selecting a random set.** Once the have formulated the available statistical information in terms of an inferential model, the next step is to select an appropriate random set on the set of all values of $U$ – i.e., a probability distribution on the class of all subsets of the range of $U$.

In the main version, the following family of sets is selected:

$$S(U) \stackrel{\text{def}}{=} \{u : |u - 0.5| \le |U - 0.5|\},$$

where $U$ is uniformly distributed on the interval $[0, 1]$. This set depends only on the value $|U - 0.5|$; so, since $|(1 - U) - 0.5| = |0.5 - U| = |U - 0.5|$, we conclude that $S(U) = S(1 - U)$. Thus, it is sufficient to describe such sets for $U \in [0.5, 1]$.

Each such value can be described as $U = 0.5 + \beta/2$, where $\beta \in [0, 1]$ is uniformly distributed on the interval $[0, 1]$. The corresponding set $S(U)$ takes the form

$$S_\beta \stackrel{\text{def}}{=} \left[\frac{1}{2} - \frac{\beta}{2}, \frac{1}{2} + \frac{\beta}{2}\right].$$

In the following text, this is the form that we will use.

**Third step of the inferential model approach.** On the third step, for each value $u$, we define $\Theta_s(u) \stackrel{\text{def}}{=} \{\theta : s = a(\theta, u)\}$ and then, for each set $S$, we define $\Theta_s(S) \stackrel{\text{def}}{=} \bigcup_{u \in S} \Theta_s(u)$.

In our example, $\Theta_s(u)$=$\{\theta : s = G_\theta^{-1}(u)\}$, i.e.,

$$\Theta_s(u) = \{\theta : G_\theta(s) = u\}.$$

Correspondingly, we have $\Theta_S(u) = \{\theta : G_\theta(s) \in S\}$. So, for sets

$$S_\beta = \left[\frac{1}{2} - \frac{\beta}{2}, \frac{1}{2} + \frac{\beta}{2}\right],$$

we have

$$\Theta_S(u) = \left\{\theta : \frac{1}{2} - \frac{\beta}{2} \le G_\theta(s) \le \frac{1}{2} + \frac{\beta}{2}\right\}.$$

**Fourth step of the inferential models approach: computing the plausibility function.** On the fourth step, we compute the plausibility

$$\text{pl}_s(\theta) \stackrel{\text{def}}{=} \text{Prob}(\theta \in \Theta_s(S)\},$$

where the probability is taken over the random set $S(U)$.

In our case, the condition $\theta \in \Theta_s(S_\beta)$ is equivalent to $|G_\theta(s) - 0.5| \le \dfrac{\beta}{2}$, i.e., to $\beta \ge 2 \cdot |G_\theta(s) - 0.5|$. Since $\beta$ is uniformly distributed on the interval $[0, 1]$, the probability for $\beta$ to satisfy this inequality is equal to the length of the interval $[2 \cdot |G_\theta(s) - 0.5|, 1]$ formed by all values $\beta$ that satisfy this inequality. So,

$$\text{pl}_s(\theta)\text{=Prob}(\theta \in \Theta_s(S)) = 1 - 2 \cdot |G_\theta(s) - 0.5|.$$

**Final step of the inferential models approach: designing the confidence regions.** According to the inferential models approach, for each $\alpha$ from the interval $(0, 1)$, we select the region $\{\theta : \text{pl}_s(\theta) \ge \alpha\}$.

For the above specific expression for plausibility, the inequality $\text{pl}_s(\theta) \ge \alpha$ takes the form $1 - 2 \cdot |G_\theta(s) - 0.5| \ge \alpha$. This inequality is equivalent to $1 - \alpha \ge |G_\theta(s) - 0.5|$, i.e., to

$$0.5 - \frac{\alpha}{2} \ge |G_\theta(s) - 0.5|.$$

An absolute value $|z|$ of any number is equal to $\max(z, -z)$. Thus, the requirement

$$0.5 - \frac{\alpha}{2} \ge |z|$$

is equivalent to requiring that

$$0.5 - \frac{\alpha}{2} \ge z \text{ and } 0.5 - \frac{\alpha}{2} \ge -z.$$

From

$$0.5 - \frac{\alpha}{2} \geq G_\theta(s) - 0.5,$$

we get $G_\theta^{-1}(s) \leq 1 - \frac{\alpha}{2}$. From

$$0.5 - \frac{\alpha}{2} \geq 0.5 - G_\theta(s),$$

we get $\frac{\alpha}{2} \leq G_\theta(s)$. Thus, the condition $\mathrm{pl}_s(\theta) \geq \alpha$ is equivalent to the double inequality

$$\frac{\alpha}{2} \leq G_\theta(s) \leq 1 - \frac{\alpha}{2}.$$

So, the inferential models approach leads to following confidence region.

**Resulting confidence regions.** According to the main version of the inferential models approach, for each $\alpha$, we select the following confidence region:

$$C(s) = \left\{\theta : \frac{\alpha}{2} \leq G_\theta(s) \leq 1 - \frac{\alpha}{2}\right\}.$$

# 3 A Simplified Way to Derive the Corresponding Confidence Regions

Let us show that the confidence regions designed in the main version of the inferential models approach can be derived in a much simpler way, without the need to invoke random sets.

Indeed, for each $\theta \in \Theta$, based on the following facts:

- that each $x_i$ is distributed according to the distribution $f_\theta(x_i)$ and
- that different $x_i$ are independent random variables,

we can determine the resulting distribution for $s(x)$. Let us denote the corresponding cumulative distribution function by $G_\theta(t)$. The probability distribution $G_\theta$ describes, for each $\theta$, the probabilities that the statistic $s(x)$ takes different values.

In particular, for each $\theta$, the probability that $s(x)$ is smaller than or equal to $G_\theta^{-1}\left(\frac{\alpha}{2}\right)$ – i.e., equivalently, that $\frac{\alpha}{2} \leq G_\theta(s(x))$ – is equal to $\alpha/2$. Similarly, the probability that $s(x)$ is grater than or equal to $G_\theta^{-1}\left(1 - \frac{\alpha}{2}\right)$ – i.e., equivalently, that $G_\theta(s(x)) \leq 1 - \frac{\alpha}{2}$ – is also equal to $\frac{\alpha}{2}$.

Thus, for every $\theta$,

$$\mathrm{Prob}\left(\frac{\alpha}{2} \leq G_\theta(s(x)) \leq 1 - \frac{\alpha}{2}\right) = 1 - \alpha.$$

Thus, as the desired confidence region, we can take the set

$$C(x) = \left\{\theta : \frac{\alpha}{2} \leq G_\theta(s(x)) \leq 1 - \frac{\alpha}{2}\right\}.$$

This is exactly what the main version of the inferential models approach is proposing.

## Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

The author is thankful to Hung T. Nguyen for valuable suggestions.

## References

[1] R. Martin, "Random sets and exact confidence regions", *Sankhyā: The Indian Journal of Statistics*, 2014, Vol. 76-A, Part 2, pp. 288–304.

[2] R. Martin and C. Liu, "Inferential models: a framework for prior-free posterior probabilistic inference", *Journal of American Statistical Association*, 2013, Vol. 108 (501), pp. 301–313.

[3] R. Martin and C. Liu, *Inferential Models: Reasoning with Uncertainty*, Chapman and Hall/CRC Press, Boca Raton, Florida, 2016.

[4] H. T. Nguyen, *An Introduction to Random Sets*, Chapman and Hall/CRC Press, Boca Raton, Florida, 2006.

[5] H. T. Nguyen and G. S. Rogers, *Fundamentals of Mathematical Statistics*, Springer Verlag, New York, 1989.