Data Science : Notes on confidence intervals

Confidence intervals

Some theory

There is no formal definition for a confidence interval in the lecture notes. So, in this section you will find the definition for a confidence interval in the frequentist and in the Bayesian setting.

Bayesian setting

We start with the Bayesian setting, since it is simpler. The parameter of interest Θ is a random variable. Therefore, we provide the following definition for *Bayesian confidence intervals*

Definition: Given $\alpha \in (0,1)$, a $100\alpha\%$ confidence interval for parameter $\theta \in \Theta \subset \mathbb{R}$ is any pair of functions $\ell(x)$ and $h(x)$, such that

$$\mathbb{P}(\ell(x) \leq \Theta \leq h(x) | X = x) \geq \alpha$$

Note: In the Bayesian setting, confidence intervals are called credible intervals.

Note 2: There might be more than one confidence intervals for the same distribution. For example, in Figure (i) and Figure (ii) below the boundaries of the interval are different but the probability that θ is in the interval is the same (equal to $1-\alpha$). Sometimes, we prefer some confidence intervals over others. For example, for most distributions we prefer the shortest possible confidence intervals. For the normal distribution in Figure (iii), this is also symmetric around μ.

Frequentist setting

In the frequentist setting, the parameter θ of the distribution is not a random variable (it is fixed), so $\mathbb{P}(\theta|X)$ makes no sense. So, how should we interpret a confidence interval? The only random variable that we have is the data X that is generated using this fixed θ. So, the estimated interval will be a function of the observed data x, let's say $[\ell(x), h(x)]$ (with $\ell(x) \leq h(x)$).

Definition: Given $\alpha \in (0,1)$, a $100\alpha\%$ confidence interval for parameter $\theta \in \Theta \subset \mathbb{R}$ is any pair of functions $\ell(x)$ and $h(x)$, such that for every $\theta \in \Theta$,

$$\mathbb{P}(\ell(X) \leq \theta \leq h(X)) \geq \alpha$$

Note: The random variables here are the boundaries of the interval. This is why this procedure belongs to the theory of *interval estimation*.

Frequentist confidence intervals using resampling

Notation:

- X : the given samples from the distribution
- θ : the parameter we want to estimate

– T the function to apply to a random sample in order to get an estimate for θ
– X^* : a collection of resampled samples

Procedure:

1. Estimate $\hat{\theta} = T(X)$.
2. Resample $X_1^*, X_2^*, \ldots, X_k^*$ from X.
3. For each collection of resampled samples, estimate the parameter θ, to obtain $\hat{\theta_1} = T(X_1^*), \ldots, \hat{\theta_k} = T(X_k^*)$.
4. Estimate the empirical distribution $\delta_1 = \hat{\theta_1} - \hat{\theta}, \ldots, \delta_k = \hat{\theta_k} - \hat{\theta}$.
5. By sorting δs, find ℓ and h such that $\hat{\mathbb{P}}(\ell \leq \delta \leq h) = a$. Usually we find the point ℓ such that $\hat{\mathbb{P}}(\delta \leq \ell) = \frac{1-\alpha}{2}$ (and the respective point for h).

Assumptions:

1. The point estimate $\hat{\theta} = T(X)$ of the parameter is close to the true value θ.
2. We assume that $T(X^*) - \hat{\theta}$ approximates well $\hat{\theta} - \theta$.

Given these two assumptions, we found ℓ and h such that

$$\mathbb{P}(\ell \leq T(X^*) - \hat{\theta} \leq h) = a$$

Since $\mathbb{P}(\ell \leq T(X^*) - \hat{\theta} \leq h) \approx \mathbb{P}(\ell \leq \hat{\theta} - \theta \leq h)$, we get

$$\mathbb{P}(\ell \leq T(X) - \theta \leq h) = a \Rightarrow$$
$$\mathbb{P}(T(X) - h \leq \theta \leq T(X) - \ell) = a$$

which gives the confidence interval for θ.

Revision

– Define a confidence interval in a frequentist setting.
– Define a confidence interval in a Bayesian setting.
– Explain the difference between the two.
– Describe how to estimate a confidence interval using resampling in a frequentist setting.
– Explain how this is applied to Question 3 Example Sheet 3.
– Describe how to estimate a confidence interval by sampling models in a Bayesian setting. Write down the assumptions and procedure.
– Explain how this is applied to Question 4 Example Sheet 2.

Exercises

Exercise [Normal confidence interval]: Derive a confidence interval for the mean of the normal distribution (with known variance) given n samples. Why do we usually choose confidence intervals of equal tails?

Exercise: (+) Is it true that for a distribution that is symmetric about x_0, the shortest possible confidence interval is symmetric about x_0? Explain why.

Exercise [Uniform distribution]: In this exercise, you will construct (meaningful) confidence intervals for the parameter of the uniform distribution. You are given i.i.d. r.vs. $X_1, \ldots, X_n \sim U[0, \theta]$.

1. Consider the r.v. $Q = \frac{1}{\theta} \max_{i=1,\dots,n} X_i$, show that $\mathbb{P}(Q \leq t) = t^n$ (see order statistics exercises).
2. Show that $\mathbb{P}(Q \leq 1) = 1$.
3. By considering $\mathbb{P}(t \leq Q \leq 1)$, construct a $1 - \alpha$ confidence interval.