NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

Lecture 12 - Confidence intervals. Introduction to Hypothesis testing.

In practice we often want, instead of a 'point estimate' for ϑ or a function $f(\vartheta)$, to give an interval $[\Theta^+, \Theta^-]$ such that $f(\vartheta) \in [\Theta^+, \Theta^-]$ 'with a high degree of confidence".

Definition 1 (Confidence interval) *Let $0 < \alpha < 1$.*[1] *Given a parametric model,*[2] *a pair of statistics $\widehat{\Theta}^-$ and $\widehat{\Theta}^+$, with $\widehat{\Theta}^- \leq \widehat{\Theta}^+$ describe a $1-\alpha$* confidence interval *for a function $f(\vartheta)$ if*

$$\mathbb{P}_\vartheta(\widehat{\Theta}^- \leq f(\vartheta) \leq \widehat{\Theta}^+) \geq 1-\alpha$$

for every $\vartheta \in \Theta$.

An interpretation: We cannot make a statement like: "$f(\vartheta) \in [\widehat{\Theta}^-, \widehat{\Theta}^+]$ with probability 95%". This simply does not make any sense, since there is no underlying probability space. Instead, what we want to say is: "Whatever the true probability measure is, our estimated interval $[\widehat{\Theta}^-, \widehat{\Theta}^+]$ will contain $f(\vartheta)$ at least 95% of the time we sample."

Example 1 *Assume $X \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known and $\vartheta = \mu$ is unknown (e.g. we measure the room temperature μ, and σ^2 is a known property of the thermometer). Fix $\alpha \in (0,1)$. Put*

$$z_{\alpha/2} = \Phi^{-1}(1-\alpha/2),$$

where, as usual, Φ is the cdf of $\mathcal{N}(0,1)$ and Φ^{-1} is its (uniquely defined) inverse function: $\Phi(\Phi^{-1}(x)) = \Phi^{-1}(\Phi(x)) = x$ for all $x \in \mathbb{R}$. Note that $1-\alpha/2 \geq 1/2$, implying $z_{\alpha/2} > 0$. Now we define

$$\widehat{\Theta}^- = \bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad \widehat{\Theta}^+ = \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

and claim that $[\widehat{\Theta}^-, \widehat{\Theta}^+]$ is a $1-\alpha$ confidence interval for μ. Indeed,

$$\mathbb{P}_\mu\left(|\bar{X}_n - \mu| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}_\mu\left(\left|\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}\right).$$

Noting that

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}),$$

and $Y = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, we obtain

$$\mathbb{P}_\mu\left(\left|\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}\right) = \mathbb{P}(-z_{\alpha/2} \leq Y \leq z_{\alpha/2}) = \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}).$$

[1] Typically α is small. Very often we shall use $\alpha = 0.05$.

[2] This can be extended to non-parametric models in the usual way

The symmetry of the normal distribution implies $\Phi(x) + \Phi(-x) = 1$ for all x, and therefore

$$\Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 2\Phi(z_{\alpha/2}) - 1 = 2(1 - \alpha/2) - 1 = 1 - \alpha.$$

What if X is no longer assumed to be normally distributed, and, as previously, $\sigma = \mathrm{Var}(X)$ is known and $\mu = \mathbb{E}(X)$ is unknown? For large n the same interval as above will yield asymptotically a $1 - \alpha$ confidence interval for μ. This is because of CLT:

$$\lim_{n\to\infty} \mathbb{P}_\vartheta(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha.$$

Now, how to design a confidence interval for μ in the normal model if both μ and σ are unknown?

Fact 1 (Student's t-distribution) *If $X_1, \dots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ variables then*

$$Y_n = \frac{\bar{X}_n - \mu}{\sqrt{\hat{S}_n^2}/\sqrt{n}}$$

is distributed with the so-called Student's t-distribution with $n - 1$ degrees of freedom, *characterized by the pdf*

$$f_{Y_n}(y) = \frac{\Gamma(\frac{n}{2})}{\sqrt{(n+1)\pi}\Gamma(\frac{n-1}{2})} \left(1 + \frac{y^2}{n-1}\right)^{-n/2}.$$

Its cdf is denoted Ψ_{n-1}, and is numerically well-understood (tables). Moreover, we have $\mathbb{E}(Y_n) = 0$,

$$\mathrm{Var}(Y_n) = \begin{cases} \frac{n-1}{n-3}, & n \geq 4 \\ \infty, & n \leq 3 \end{cases}$$

and $Y_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Example 2 *Assume $X \sim \mathcal{N}(\mu, \sigma^2)$, where both μ and σ are unknown, and let $\alpha \in (0, 1)$. Put*

$$t_{\alpha/2} = \Psi_{n-1}^{-1}(1 - \alpha/2),$$

and for $n \geq 4$ samples we set

$$\widehat{\Theta}^- = \bar{X}_n - t_{\alpha/2}\frac{\sqrt{\hat{S}_n^2}}{\sqrt{n}}, \qquad \widehat{\Theta}^+ = \bar{X}_n + t_{\alpha/2}\frac{\sqrt{\hat{S}_n^2}}{\sqrt{n}}.$$

This gives a $1 - \alpha$ confidence interval for μ. The proof is the same as in Example 1, with Φ replaced everywhere by Ψ (due to Fact 1).

Note that, because of $Y_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$, for very large n we would not be too wrong also using Φ here.

Let us now study the confidence intervals in the non-parametric setting. Suppose that all we know about the distribution of X is that it is continuous, and in particular,

$$\mathbb{P}(\bigcup_{i\neq j}\{X_i = X_j\}) = 0,$$

in other words, no two values of the samples will coincide, a.s. In this situation we cannot expect to produce a reliable confidence interval for the mean, as the distribution could have 'heavy tails', i.e., deviate from the mean in both directions, with a large probability. We can, however, give a confidence interval for the *median*, that is, the value $m = F_X^{-1}(1/2)$ (satisfying $\mathbb{P}(X > m) = \mathbb{P}(X < m) = 1/2$). To do so, we use an *order statistic*.

Definition 2 (Order statistic) *The* k-th order statistic *of the statistical sample* $X_1, \ldots, X_n$ *is the* k*-th smallest value among* $X_1, \ldots, X_n$*, and is denoted by* $X_{(k)}$.

In particular, $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Theorem 1 *Given* n *and* $0 < \alpha < 1$*, and let* k *be the largest integer with*

$$F_Y(k-1) \leq \alpha/2,$$

where $Y \sim Bin(n, 1/2)$*. Then* $[X_{(k)}, X_{(n-k+1)}]$ *defines a* $1-\alpha$ *confidence interval for the median* $m = m(\mathbb{P})$.

Proof For every $\mathbb{P} \in \mathcal{P}$ and every $1 \leq i \leq i$ we have $\mathbb{P}(X_i \leq m) = 1/2$, and these events are independent. Therefore, $Y = \sum_{i=1}^n \mathbb{1}_{\{X_i \leq m\}}$ is $Bin(n, 1/2)$-distributed, and we obtain

$$\mathbb{P}(m < X_{(k)}) = \mathbb{P}(Y \leq k-1) = F_Y(k-1) \leq \alpha/2.$$

And, by symmetry of the binomial distribution,

$$\begin{aligned}\mathbb{P}(m > X_{(n-k+1)}) &= \mathbb{P}(Y \geq n-k+1) = \mathbb{P}(Y \leq k-1) \\ &= F_Y(k-1) \leq \alpha/2.\end{aligned}$$

In total,

$$\begin{aligned}\mathbb{P}(m \in [X_{(k)}, X_{(n-k+1)}]) &= 1 - \mathbb{P}(m < X_{(k)}) - \mathbb{P}(m > X_{(n-k+1)}) \\ &\geq 1 - \alpha/2 - \alpha/2 = 1 - \alpha.\end{aligned}$$

□

Remark 1 *One can similarly determine confidence intervals for* $F_X(t)$*, for any fixed* $0 < t < 1$.

Let us now turn our attention to a third method of statistical inference, namely Hypothesis testing.

Suppose we are investigating the bias of a coin. Quite often we have a suspicion, or a specific question we want to resolve/answer in a binary form. For example, we may want to decide if our coin is fair or not fair. Alternatively, if the coin is head-biased or tail-biased. Formally, we consider a statistical model $(\Omega, \mathcal{F}, \mathbb{P}_\vartheta : \vartheta \in \Theta)$, alongside a partition $\Theta = \Theta_0 \cup \Theta_1$ (disjoint subsets). The sets Θ_0 and Θ_1 are associated with the so-called *null hypothesis* (typically reflecting one's default assumption) and the *alternative*, respectively. We aim to decide whether H_0 or H_1 holds (accept or reject the null-hypothesis). Doing so we might commit two types of error.

- type I error: false rejection. H_0 was true, but we rejected it.
- type II error: false acceptance. H_0 was false, but we accepted it.

As before, we make our decision based on a sample of n iid variables $X_1, \dots, X_n \sim \mathbb{P}_\vartheta$. A decision rule involves choosing

- The *test statistic* $S = h(X_1, \dots, X_n)$, where $h : \mathbb{R}^n \to \mathbb{R}$ is a function.
- The *rejection region* $W \subseteq \mathbb{R}$.

We sample the values $x_1, \dots, x_n$ of $X_1, \dots, X_n$ and apply the decision rule: reject H_0 if $h(x_1, \dots, x_n) \in W$ and accept H_0 otherwise.

The *significance level* of the test is defined as $1 - \alpha$, where

$$\alpha = \sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta(S \in W)$$

That is, α is the maximal[3] probability of a type I error. [4] The value α is typically set in advance as a requirement for the test.

The *power* of the test is defined as $1 - \beta(\vartheta)$, where $\beta : \Theta_1 \to [0, 1]$ is the function

$$\beta(\vartheta) = \mathbb{P}_\vartheta(S \notin W).$$

That is, $\beta(\vartheta)$ is the probability of a type II error, viewed as a function of $\vartheta \in \Theta_1$.

Our goal is to design a test at significance level (at most) $1 - \alpha$, while minimizing the value(s) of β.

Remark 2 *In practice, to avoid statistical malpractice, it is crucial that the decision rule be clearly formulated* ***before*** *conducting the sampling.*

[3]supremal to be precise, but in practice the maximum is usually attained

[4]A very popular value is $\alpha = 0.05$