BOOTSTRAP CONFIDENCE INTERVALS [INTEGRATION TD2]

Setting [based on Exercise 2]

We have a sample of $N = 2^n$ data $(x_1, \cdots, x_N)$ generated from an unknown population of size M (more generally, from a distribution with density f). We take some "statistics" (=function of the data) $T(x_1, \cdots, x_N)$, in the exercise $T(x_1, \cdots, x_n) = \min_i x_i$. We denote with $\hat{t}$ the value of $T(x_1, \cdots, x_n)$ on the particular sample that we have. Notice that there is a "true" value of T, that is the minimum over the population of size M. In the exercise we assume that the sample has a simple form, $x_i \in \{2^j\}_{j=1}^n$, and 2^j appears with multiplicity $m(j)$. In this case our sample gives $\hat{t} = \min_i x_i = 2$.

The estimate $\hat{t}$ is a random variable itself, which depends on the sample [if I change the sample, the corresponding value of $\hat{t}$ changes]; its distribution is unknown, because f is unknown. We want to use bootstrap to estimate its distribution, variance, confidence intervals.

The procedure is now:

(i) Approximate the unknown f with the empirical density $\hat{f}$ obtained from the sample: $\hat{f} = \frac{1}{N}\sum_{i=1}^N \delta(x - x_i)$, in our case:

$$\hat{f}(x) = \sum_{j=1}^{n+1} \frac{m(j)}{N}\delta(x - 2^j). \tag{1}$$

(ii) Sample "bootstrap realizations" [i.e., other sets of data $(x_1^*, \cdots, x_N^*)$] from the empirical density $\hat{f}$; each of them gives a bootstrap realization of the statistics T, which we call $t^* = \min_i x_i^*$;

(iii) compute the distribution of t^* over the bootstrap realizations, the moments [see Ex. 2 (a)], and confidence intervals [see Ex. 2(c)].

Confidence interval and interpretation

In principle we want to find an interval $[a, b]$ such that:

$$P\left(T \in [a, b]\right) = 1 - \alpha. \tag{2}$$

However, without knowing f, the only thing that we can get is actually:

$$P\left(\hat{t} \in [a, b]\right) = 1 - \alpha, \tag{3}$$

where a, b depend on the sample. This has to be interpreted as follows [see Wasserman Sec. 6.3.2]:

- On day 1, I have a sample $x_1, \cdots, x_N$ and I compute (i) the estimate $\hat{t}$ from this sample, (ii) the the constants a, b (that depend on $\hat{t}$, see below); this gives the interval $[a_{day1}, b_{day1}]$
- On day 2, I have another sample and I get another value of $\hat{t}$ and a new interval $[a_{day2}, b_{day2}]$
- After I repeat infinitely many times, the $1-\alpha$ percent of the *intervals* that I constructed contains the true value of T (which in the case of the exercise, is the true minimum of the population of size M).

The bootstrap prescription [see Ex. 2 (c) and Sec. 8.3 in Wasserman] tells us that for each α we should set:

$$\begin{aligned} a &= 2\hat{t} - \mu_{1-\frac{\alpha}{2}} \\ b &= 2\hat{t} - \mu_{\frac{\alpha}{2}}, \end{aligned} \tag{4}$$

where $\mu_{1-\frac{\alpha}{2}}$ is defined from the bootstrap distribution as:

$$P\left(t^* \leq \mu_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}. \tag{5}$$

Justification

- given the function T, we impose:

$$\begin{aligned} P\left(T \leq a\right) &= \frac{\alpha}{2} \\ P\left(T \geq b\right) &= \frac{\alpha}{2}, \end{aligned} \tag{6}$$

  which ensures (3). These are equivalent to:

$$\begin{aligned} P\left(\hat{t} - T \geq \hat{t} - a\right) &= \frac{\alpha}{2} \\ P\left(\hat{t} - T \leq \hat{t} - b\right) &= \frac{\alpha}{2}. \end{aligned} \tag{7}$$

  In order to solve these equations for a, b, we should know the distribution of T, that is unknown.

- The idea of bootstrap is to replace the unknown distribution of T with the distribution constructed over the sample [corresponding to $f \to \hat{f}$], so that $T \to \hat{t}$; at the same time, the bootstrap realizations $\{x_i^*\}$ give different realizations t^*, that we can use to build a statistics for $\hat{t}$. Therefore in the equations above we substitute the variable with unknown distribution $\hat{t} - T$ with its bootstrap approximation $\hat{t} - T \to t^* - \hat{t}$.

- This gives

$$\begin{aligned} P\left(\hat{t} - T \geq \hat{t} - a\right) \approx P\left(t^* - \hat{t} \geq \hat{t} - a\right) &= \frac{\alpha}{2} \\ P\left(\hat{t} - T \leq \hat{t} - b\right) \approx P\left(t^* - \hat{t} \leq \hat{t} - b\right) &= \frac{\alpha}{2}. \end{aligned} \tag{8}$$

  Then

$$\begin{aligned} P\left(t^* \leq 2\hat{t} - a\right) &= 1 - \frac{\alpha}{2} \\ P\left(t^* \leq 2\hat{t} - b\right) &= \frac{\alpha}{2}, \end{aligned} \tag{9}$$

  and we can identify $\mu_{1-\frac{\alpha}{2}} = 2\hat{t} - a$, $\mu_{\frac{\alpha}{2}} = 2\hat{t} - b$.