18.650: Fundamentals of Statistics	October 11, 2023

Lecture 12— Intro to Bootstrap

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

The bootstrap gives us a different way to compute the variance of an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ and to construct confidence intervals, without relying on asymptotic normality.

Motivation. So far, our confidence intervals have taken the form

$$\theta \in \hat{\theta} \pm 1.96\hat{\text{se}}.$$

Note that...

- The 1.96 is based on $\hat{\theta}$ being approximately normal. This may not always be satisfied.
- So far, our estimate $\hat{\text{se}}$ of the standard error took the form $\hat{\text{se}} = \hat{\sigma}/\sqrt{n}$, using an asymptotic variance calculation based on...
  - the CLT (if $\hat{\theta} = \bar{X}_n$)
  - the Delta method (if $\hat{\theta} = g(\bar{X}_n)$)
  - the Fisher information (if $\hat{\theta} = \hat{\theta}^{\text{MLE}}$).

  However, $\hat{\theta}$ is not always one of these three forms!

Example.

Suppose we're interested in the median θ of the distribution Poisson(λ). We can formally define θ as the integer such that

$$\sum_{k=0}^{\theta-1} e^{-\lambda}\frac{\lambda^k}{k!} < 0.5, \qquad \sum_{k=0}^{\theta} e^{-\lambda}\frac{\lambda^k}{k!} \geq 0.5. \tag{1}$$

In other words, $\theta = \text{Median}\left(\text{Poisson}(\lambda)\right) = g(\lambda)$, where g is defined implicitly by (1). Now, how can we estimate θ given $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$? We could take $\hat{\theta} = g(\bar{X}_n)$, since $\bar{X}_n$ is an estimator for the mean λ. To get $\hat{\text{se}}$, we could try to apply the Delta method. But g is not differentiable — it can't even be written in closed form! A more natural solution is to take

$$\hat{\theta} = \text{Median}\left(X_1, \dots, X_n\right).$$

But now the issue is that $\hat{\theta}$ is not in one of the above three forms.

The example shows we need another way to compute standard errors.

1 The bootstrap

The setting is that we have n samples $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and we have computed an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$. We're interested in the variance of $\hat{\theta}$. Typically, a good way to get a sense of the variability of a random variable is to look at a few samples of it. But the issue is we've only observed a single $\hat{\theta}$! Although we have n samples from $\mathbb{P}$, we used up all of them to produce a *single* sample of $\hat{\theta}$. This motivates the following

1.1 Thought experiment

Suppose we could easily generate as many samples X_i as we want (in reality, these samples may be very expensive to collect). Then we could generate multiple sets of n samples, and use each set to construct a new sample of $\hat{\theta}$:

$$\begin{aligned} X_{1:n}^{(1)} &= \{X_1^{(1)}, \dots, X_n^{(1)}\} \to \hat{\theta}^{(1)} \\ X_{1:n}^{(2)} &= \{X_1^{(2)}, \dots, X_n^{(2)}\} \to \hat{\theta}^{(2)} \\ &\vdots \\ X_{1:n}^{(B)} &= \{X_1^{(B)}, \dots, X_n^{(B)}\} \to \hat{\theta}^{(B)}, \end{aligned} \tag{2}$$

where

$$X_i^{(b)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}, \quad i = 1, \dots, n, \quad b = 1, \dots, B. \tag{3}$$

We could now construct a histogram of these sample values $\hat{\theta}^{(b)}, b = 1, \dots, B$ to get a sense of their distribution. In particular, we can easily get an estimate for the variance of $\hat{\theta}$:

$$\widehat{\mathbb{V}[\hat{\theta}]} = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \frac{1}{B} \sum_{c=1}^{B} \hat{\theta}^{(c)} \right)^2 \tag{4}$$

1.2 An alternative to sampling from $\mathbb{P}$

The issue is that the sampling in (3) is too expensive, or impossible. We only have our n initial samples. So instead, we sample from

$$\hat{\mathbb{P}}_n = \text{Uniform}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

The second representation is just an alternative way to represent this uniform distribution, which is known as the *empirical distribution* of the data. The cdf corresponding to $\hat{\mathbb{P}}_n$ is known as the *empirical cdf*, and it is given by

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \le x).$$

Note that $\hat{F}_n(x)$ is actually random, since the X_i's are random. One can show by the LLN that

$$\hat{F}_n(x) \xrightarrow{\mathbb{P}} F(x), \quad n \to \infty.$$

Therefore, $\hat{F}_n$ is a good approximation to F when n is large.

Definition 1.1: Bootstrap sample

Given a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, a *bootstrap sample* is a collection of m random variables $X_1^*, \ldots, X_m^*$ such that

$$X_i^* \overset{\text{i.i.d.}}{\sim} \text{Unif}(X_1, \ldots, X_n), \quad i = 1, \ldots, m.$$

Remark.

> The bootstrap sample need not be the same size as the original sample, i.e. we could have $m \neq n$.

Example.

> Suppose we observe
>
> $$X_1 = 2.1, \quad X_2 = -1.3, \quad X_3 = 6.0, \quad X_4 = 0.7.$$
>
> The following is an example of a bootstrap sample of size $m = 5$:
>
> $$X_1^* = 6.0, \quad X_2^* = -1.3, \quad X_3^* = 6.0, \quad X_4^* = 2.1, \quad X_5^* = -1.3$$

1.3 Bootstrap variance estimation

Now that we have an alternative to sampling from $\mathbb{P}$, we can return to the scheme (2). We create B bootstrap samples,

$$\begin{aligned}
X_{1:n}^{(1)} &= \{X_1^{(1)}, \ldots, X_n^{(1)}\} \to \hat{\theta}^{(1)} \\
X_{1:n}^{(2)} &= \{X_1^{(2)}, \ldots, X_n^{(2)}\} \to \hat{\theta}^{(2)} \\
&\vdots \\
X_{1:n}^{(B)} &= \{X_1^{(B)}, \ldots, X_n^{(B)}\} \to \hat{\theta}^{(B)}
\end{aligned} \tag{5}$$

as before, but now

$$X_i^{(b)} \overset{\text{i.i.d.}}{\sim} \hat{\mathbb{P}} = \text{Unif}(X_1, \ldots, X_n), \quad i = 1, \ldots, n, \quad b = 1, \ldots, B. \tag{6}$$

As in the thought experiment, we use the sample variance as our estimator:

$$v_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \frac{1}{B} \sum_{c=1}^{B} \hat{\theta}^{(c)} \right)^2 \tag{7}$$

Note that there are two approximations in this procedure:

$$v_{\text{boot}} \approx \mathbb{V}_{\hat{\mathbb{P}}_n}[\hat{\theta}] \approx \mathbb{V}_{\mathbb{P}}[\hat{\theta}].$$

The first approximation can be made as accurate as one wants, by taking B large enough. This is not hard, because generating new bootstrap samples is cheap! The second approximation is the main limitation of the bootstrap. We need n to be large for this approximation to be good, but of course, we cannot artificially increase how many real samples we have at our disposal.