Lecture 12

23. The Bootstrap

We start the discussion about the bootstrap with an example. Suppose $X = (X_1, \dots, X_n)$ are IID unknown distribution function F. Let the empirical distribution based on X be

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

We want to construct a symmetric confidence interval for some functional of the distribution. We write it $h(F)$. For instance, if we are estimating the mean then $h(F) = \int xF(dx)$. An estimate of $h(F)$ is $h(\hat{F}_n)$ (in the case of the mean, this is the empirical mean $\bar{X}_n$) and our objective is to find t such that

$$\Pr(h(\hat{F}_n) - t \leq h(F) \leq h(\hat{F}_n) + t) = \gamma.$$

Then $I = [h(\hat{F}_n) - t, h(\hat{F}_n) + t]$ is a coefficient γ confidence interval for $h(F)$. We can write the above as

$$\Pr(-t \leq h(F) - h(\hat{F}_n) \leq t) \geq \gamma.$$

If the distribution of $h(F) - h(\hat{F}_n)$ is known then we can find out what t is. For instance, if $X_1, \dots, X_n$ are IID $N(\mu, 1)$ with μ unknown and $h(F) = \int xF(dx) = \mu$ then $h(F) - h(\hat{F}_n) = \mu - n^{-1} \sum_{i=1}^{n} X_i \sim N(0, 1/n)$ and t should be taken as the $(1+\gamma)/2$-quantile of the $N(0, 1/n)$-distribution.

The problem occurs when the distribution of $h(F) - h(\hat{F}_n)$ is unknown or too difficult to compute. Then the bootstrap methodology proposes to do the following. For each $i = 1, \dots, N$,

- Let $X^{*,i} = (X^*_{i1}, \dots, X^*_{in})$ be an independent sample from the set $\{X_1, \dots, X_n\}$ drawn uniformly with replacement. That is, $X^*_{i1}, \dots, X^*_{in}$ are IID with $\Pr(X^*_{ij} = X_k) = 1/n$. Let F^*_i be the empirical distribution of $X^*_{i1}, \dots, X^*_{in}$.
- Form $R_i = h(\hat{F}_n) - h(F^*_i)$.
- Copmute the empirical $(1+\gamma)/2$-quantile from the histogram of $R_1, \dots, R_N$ and use this for $\hat{t}$.
- Construct the confidence interval $I = [h(\hat{F}_n) - \hat{t}, h(\hat{F}_n) + \hat{t}]$.

The idea is that the distribution of $h(F) - h(\hat{F}_n)$ is unknown but we can approximate it. We approximate F by $\hat{F}_n$ and $h(F) - h(\hat{F}_n)$ by $h(\hat{F}_n) - h(F^*)$. We may not know the distribution of the latter but in any case we can simulate from it and use the simulated histogram to construct an 'approximate' confidence interval. The success of the procedure depends on to which extent the approximation of F by $\hat{F}_n$ is a good one. This can be hard to quantify.

The general setup. In general the bootstrap is set up as follow. We have $X = (X_1, \dots, X_n)$ which is an IID sample from an unknown distribution F_0. The distribution F_0 is approximated by a distribution F_1 that depends on the sample (for instance the empirical distribution). Given a functional f_t from a class $\{f_t : t \in \mathcal{T}\}$ we wish to find a value t_0 that solves

$$E[f_t(F_0, F_1) \mid F_0] = 0. \tag{23.1}$$

This is called the population equation. For example if $h(F_0) = (\int xF_0(dx))^r$ and $F_1 = \hat{F}_n$, then $h(F_1) = (n^{-1}\sum_{i=1}^n X_i)^r$. This estimate of $h(F_0)$ will typically be biased. To correct for the bias we can take

$$f_t(F_0, F_1) = h(F_1) - h(F_0) + t \tag{23.2}$$

and t_0 that solves (23.1). The bias corrected estimate is $h(F_1) + t_0$. For the symmetric confidence interval we would take

$$f_t(F_0, F_1) = I\{h(F_1) - t \leq h(F_0) \leq h(F_1) + t\} - \gamma. \tag{23.3}$$

In some situations, such as with the confidence interval for the normal sample above, we can solve the sample equation. However, in many situations we cannot. Then we can try to obtain as approximate solution by replacing F_0 by F_1 and F_1 by F_2 where F_2 is a distribution that depend on a sample drawn from F_1. The resulting equation is then

$$E[f_t(F_1, F_2) \mid F_1] = 0.$$

This is called the sample equation. The solution $\hat{t}_0$ to the sample equation is used instead of t_0. The idea is that $\hat{t}_0$ is a good approximation of t_0. This is called the *bootstrap principle*.

There are essentially two ways to choose F_1 and F_2.

- F_1 is the empirical distribution $\hat{F}_n$ of $X = (X_1, \ldots, X_n)$. This is referred to as the nonparamteric bootstrap. F_2 is then taken to be the empirical distribution F^* of an IID sample $X^* = (X_1^*, \ldots, X_n^*)$ from $\hat{F}_n$.
- If F_0 is assumed to belong to a parametric family $\mathcal{F} = \{F_\theta : \theta \in \Omega\}$. Let $\hat{\Theta}$ be an estimate of Θ. Then we take $F_1 = F_{\hat{\Theta}}$. Let $X^* = (X_1^*, \ldots, X_n^*)$ be an IID sample from F_1 and $\hat{\Theta}^*$ an estimate of Θ based on X^*. Then we put $F_2 = F_{\hat{\Theta}^*}$. This is called the parametric bootstrap.

Bias reduction. Let f_t be given by (23.2). Then the sample equation assumes the form

$$E[h(F_2) - h(F_1) + t \mid F_1] = 0$$

and the solution is

$$\hat{t}_0 = h(F_1) - E[h(F_2) \mid F_1].$$

Thus the bootstrap-reduced bias estimate is

$$h(F_1) + \hat{t}_0 = 2h(F_1) - E[h(F_2) \mid F_1].$$

The expected value $E[h(F_2) \mid F_1]$ may be difficult to compute analytically but can always be computed by simulation: For $i = 1, \ldots, N$

- Let $X^{*,i} = (X_{i1}^*, \ldots, X_{in}^*)$ be an IID sample from the set $\{X_1, \ldots, X_n\}$ drawn from F_1 ($F_1 = \hat{F}_n$ in nonparametric case and $F_1 = F_{\hat{\Theta}}$ in parametric case) . Let F_i^* be the empirical distribution of $X_{i1}^*, \ldots, X_{in}^*$ in nonparametric case and $F_i^* = F_{\hat{\Theta}_i^*}$ where $\hat{\Theta}_i^*$ is estimate of Θ based on $X^{*,i}$ in parametric case.
- Put $R_i = h(F_i^*)$.
- Compute $E[h(F_2) \mid F_1]$ by $N^{-1}\sum_{i=1}^N R_i$.

We can get arbitrary accuracy of $E[h(F_2) \mid F_1]$ by taking N sufficiently large (by the SLLN). However, the performance depend on how good the approximation F_1 of F_0 is. (Hall, 1992, "The Bootstrap and Edgeworth Expansions" obtain asymptotic results that indicate the accuracy of the approximations).

Example 32. Let $\mu = \int xF(dx)$ and suppose we want to estimate μ^3. We put $h(F_0) = \mu^3$. In the nonparametric case, with F_1 being the empirical distribution $\hat{F}_n$ we have $h(F_1) = \bar{X}_n^3$ (with $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$). We can calculate

$$\begin{aligned} E[h(F_1) \mid F_0] &= E\Big[\Big(\mu + n^{-1}\sum_{i=1}^n (X_i - \mu)\Big)^3\Big] \\ &= \mu^3 + n^{-1}3\mu\sigma^2 + n^{-2}\gamma \end{aligned}$$

with $\sigma^2 = E[(X_1 - \mu)^2]$ and $\gamma = E[(X_1 - \mu)^3]$. Similarly we compute

$$E[h(F_2) \mid F_1] = \bar{X}_n^3 + n^{-1}3\bar{X}_n\hat{\sigma}^2 + n^{-2}\hat{\gamma}$$

with $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\gamma = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^3$. The bootstrap bias-reduced estimate is then

$$\begin{aligned} \hat{\Theta} &= 2h(F_1) - E[h(F_2) \mid F_1] = 2\bar{X}_n^3 - (\bar{X}_n^3 + n^{-1}3\bar{X}_n\hat{\sigma}^2 + n^{-2}\hat{\gamma}) \\ &= \bar{X}_n^3 - n^{-1}3\bar{X}_n\hat{\sigma}^2 - n^{-2}\hat{\gamma}. \end{aligned}$$

Confidence intervals. A symmetric confidence interval may be constructed using the function f_t in (23.3). Then the sample equation becomes

$$P(h(F_2) - t \le h(F_1) \le h(F_2) + t \mid F_1) - \gamma = 0. \tag{23.4}$$

Since the distribution of $h(F_2)$ conditional on F_1 is discrete we may not be able to solve with equality but if n is not very small we can come very close. In any case we could take

$$\hat{t}_0 = \inf\{t : P(h(F_2) - t \le h(F_1) \le h(F_2) + t \mid F_1) - \gamma \ge 0\}.$$

We could also consider other confidence intervals such as $(h(F_1) - \hat{t}_{01}, h(F_1) + \hat{t}_{02})$ where $\hat{t}_{01}$ and $\hat{t}_{02}$ solve

$$\begin{aligned} P(h(F_1) \le h(F_2) - t \mid F_1) - (1-\gamma)/2 = 0, \\ P(h(F_1) \le h(F_2) + t \mid F_1) - (1+\gamma)/2 = 0, \end{aligned}$$

or one-sided intervals $(-\infty, h(F_1) + \hat{t}_{03})$ where $\hat{t}_{03}$ solves

$$P(h(F_1) \le h(F_2) + t \mid F_1) - \gamma = 0.$$

To find these $\hat{t}_{01}, \hat{t}_{02}$, and $\hat{t}_{03}$ we can generate a sample of size N from $h(F_1) - h(F_2)$ and take the appropriate empirical quantiles.

Pivotal quantities. Suppose again we want to construct a symmetric confidence interval so the sample equation is given by (23.4). The performance of the bootstrap method will depend on if the distribution of $h(F_1) - h(F_2)$ is a good approximation for the distribution of $h(F_0) - h(F_1)$. In some situations we can get an approximation for the variance of this distribution. Let us consider an example where explicit computations are possible, to illustrate the point.

Let $X_1, \dots, X_n$ be IID $N(\mu, \sigma^2)$ with unknown μ and σ and $h(F_0) = \mu = \int xF(dx)$. In this case we know that with $S_n = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ the quantity

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

has a student-t distribution with $n-1$ degrees of freedom. let T_{n-1} be the cdf of this distribution. To construct the confidence interval we choose t such that

$$P(-t \leq \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \leq t) = \gamma. \tag{23.5}$$

That is $t = T_{n-1}^{-1}((1+\gamma)/2)$, the $(1+\gamma)/2$-quantile. The resulting coefficient γ confidence interval is $(\bar{X}_n - tS_n/\sqrt{n}, \bar{X}_n + tS_n/\sqrt{n})$. Obviously, in this case there is no need for the bootstrap, but for comparison let's see what the bootstrap methodology would give us in this case.

We would use the parametric bootstrap. We could use the estimates $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma}^2(F_1) = S_n^2$. Then $F_1 = F_{\hat{\mu},\hat{\sigma}^2}$ is the cdf of a $N(\bar{X}_n, \sigma^2(F_1))$ distribution. Note that $h(F_2) = \bar{X}_n^*$ where $X^* = (X_1^*, \dots, X_n^*)$ is an IID sample from F_1.

The quantity we really want is $h(F_0) - h(F_1) = \mu - \bar{X}_n \sim N(0, \sigma^2/n)$ but σ is unknown. Using the sample equation we want to find the solution to

$$P(-t \leq h(F_1) - h(F_2) \leq t \mid F_1) = \gamma.$$

Now we see that $h(F_1) - h(F_2) = \bar{X}_n - \bar{X}_n^* \sim N(0, \sigma^2(F_1)/n)$. Equivalently we can write sample equation as

$$P(-t \leq \frac{\sigma(F_1)}{\sqrt{n}} Z \leq t \mid F_1) = \gamma$$

where $Z \sim N(0,1)$. Then we see that t should be taken as $\Phi^{-1}((1+\gamma)/2)$, the $(1+\gamma)/2$-quantile of the standard normal and the resulting confidence interval is $(\bar{X}_n - tS_n/\sqrt{n}, \bar{X}_n + tS_n/\sqrt{n})$. We see that the difference from the exact confidence interval comes from the approximation of a t-quantile $T_{n-1}^{-1}((1+\gamma)/2)$ by a standard normal quantile $\Phi^{-1}((1+\gamma)/2)$.

However, if we instead of $h(F_0) - h(F_1)$ considers $[h(F_0) - h(F_1)]/\hat{\sigma}(F_1)$ where $\sigma^2(F_1)$ is an estimate of the variance of $h(F_0)$, then we will do much better (note that above we used $\hat{\sigma}^2(F_1)$ as an estimate of the variance of F_0 but here we want an estimate of the variance of $h(F_0)$ so when $h(F_0) = \mu$ we could take $\sigma^2(F_1) = S_n^2/n$). In this case we would take

$$f_t(F_0, F_1) = I\{-t \leq \frac{h(F_0) - h(F_1)}{\sigma(F_1)} \leq t\} - \gamma.$$

The population equation becomes

$$P(-t \leq \frac{h(F_0) - h(F_1)}{\sigma(F_1)} \leq t \mid F_0) = \gamma.$$

If $\sigma^2(F_1) = S_n^2/n$ then we see that this is again (23.5) and we can solve the population equation exactly. The important thing is that the ratio $\frac{h(F_0)-h(F_1)}{\hat{\sigma}(F_1)}$ has a distribution that does not depend on the parameters. It is called a *pivotal quantity*. Then the bootstrap method is also expected to work well. Suppose we didn't recognize that we can actually solve the population equation and we proceed instead

with the parametric bootstrap. Then the sample equation becomes

$$P(-t \leq \frac{h(F_1) - h(F_2)}{\sigma(F_2)} \leq t \mid F_1) = \gamma.$$

Note that

$$\frac{h(F_1) - h(F_2)}{\sigma(F_2)} = \frac{\bar{X}_n - \bar{X}_n^*}{S_n^*/\sqrt{n}}$$

given F_1 also has a student-t distribution with $n-1$ degrees of freedom. To find $\hat{t}_0$ we have to take the $(1+\gamma)/2$-quantile, i.e. $\hat{t}_0 = T_{n-1}^{-1}((1+\gamma)/2)$ and the resulting confidence interval is $(\bar{X}_n - \hat{t}_0 S_n/\sqrt{n}, \bar{X}_n + \hat{t}_0 S_n/\sqrt{n})$. This is exactly the correct coefficient γ confidence interval. The key issue here is that we could find a pivotal quantity that did not depend on the parameters. Often this method of dividing by some standard deviation estimate works well. The reason is that the quantity $\frac{h(F_0)-h(F_1)}{\sigma(F_1)}$ may not be pivotal but often close to pivotal. Its distribution is better approximated by the bootstrap method than the distribution of $h(F_0) - h(F_1)$.