Lecture Notes 20
36-705

We have discussed plug-in estimators and influence functions. Today we consider a different nonparametric approach for getting confidence intervals for plug-in estimators: the bootstrap.

1 Monte Carlo

Before we get to the bootstrap, we should briefly discuss the Monte Carlo method.

Let g be a function and let P be a distribution. Suppose we want to know the mean of g, that is $\mathbb{E}[g(X)] = \int g(x)p(x)dx$. One way to do this is to do the integral $\int g(x)p(x)dx$. Another approach is simulation, also known as Monte Carlo. We draw a large sample $X_1, \ldots, X_B \sim P$. Then, by the law of large numbers

$$\frac{1}{B}\sum_{j=1}^{B} g(X_j) \xrightarrow{P} \mathbb{E}[g(X)].$$

Since we can simulate as many observations as we want, we can make the estimate very close to $\mathbb{E}[g(X)]$.

The same is true for the variance. We can get the variance of $g(X)$ by integration:

$$\text{Var}[g(X)] = \int g^2(x)p(x) - \left(\int g(x)p(x)\right)^2.$$

But we can also compute the sample variance from the simulated values and, again, by the law of large numbers

$$\frac{1}{n}\sum_j (g(X_j) - \overline{g})^2 \xrightarrow{P} \text{Var}[g(X)]$$

where $\overline{g} = \frac{1}{B}\sum_j g(X_j)$.

Now suppose that $T = g(X_1, \ldots, X_n)$ is a function of n iid variables. The mean is

$$\int \cdots \int g(x_1, \ldots, x_n)p(x_1)\cdots p(x_n)dx_1 \ \ldots \ dx_n$$

which is an n-dimensional integral. We can still use Monte-Carlo if we draw samples of size n each time. When we draw $X_1, \ldots, X_n \sim p$, we can think of this as one draw from the joint

density $p(x_1, \ldots, x_n) = p(x_1) \cdots p(x_n)$. In other words we do the following:

$$\begin{aligned} \text{draw } X_1, \ldots, X_n \sim P \quad & \text{compute } T_1 = g(X_1, \ldots, X_n) \\ \text{draw } X_1, \ldots, X_n \sim P \quad & \text{compute } T_2 = g(X_1, \ldots, X_n) \\ & \vdots \\ \text{draw } X_1, \ldots, X_n \sim P \quad & \text{compute } T_B = g(X_1, \ldots, X_n). \end{aligned}$$

Then $T_1, T_2, \ldots$ are draws from the distribution of $T = g(X_1, \ldots, X_n)$. Again, by the law of large numbers, as $B \to \infty$,

$$\frac{1}{B} \sum_{j=1}^{B} T_j \xrightarrow{P} \mathbb{E}[T] = \mathbb{E}[g(X_1, \ldots, X_n)]$$

and

$$\frac{1}{n} \sum_j (T_j - \overline{T})^2 \xrightarrow{P} \text{Var}[T] = \text{Var}[g(X_1, \ldots, X_n)]$$

where $\overline{T} = \frac{1}{B} \sum_j T_j$.

2 Bootstrap Variance Estimation

Let $X_1, \ldots, X_n \sim P$ and let $T = g(X_1, \ldots, X_n)$ be some statistic. Of course, the case we have in mind is that $T = g(X_1, \ldots, X_n)$ is an estimator of some parameter. Our goal is to estimate the standard error, that is the standard deviation of T. As a concrete example, think of $T = g(X_1, \ldots, X_n)$ as the median of the data.

If we knew P, we could use Monte Carlo to estimate $\tau^2 = \text{Var}[T]$. The idea of the bootstrap is to estimate P with the empirical distribution P_n. In other words, τ^2 is a statistical functional so we can write it as $\tau^2(P)$. We will estimate $\tau^2(P)$ with $\tau^2(P_n)$. Computing $\tau^2(P_n)$ is not easy to do analytically but now we can use Monte Carlo. We just need to simulate many times from P_n. When we draw a sample from P_n we usually denote the draws by X_i^*. So

$$X_1^*, \ldots, X_n^* \sim P_n$$

denotes a sample from P_n. We call $X_1^*, \ldots, X_n^*$ a bootstrap sample.

Specifically:

$$\begin{aligned} \text{draw } X_1^*, \ldots, X_n^* \sim P_n \quad & \text{compute } T_1 = g(X_1^*, \ldots, X_n^*) \\ \text{draw } X_1^*, \ldots, X_n^* \sim P_n \quad & \text{compute } T_2 = g(X_1^*, \ldots, X_n^*) \\ & \vdots \\ \text{draw } X_1^*, \ldots, X_n^* \sim P_n \quad & \text{compute } T_B = g(X_1^*, \ldots, X_n^*). \end{aligned}$$

Again, by the law of large numbers, as $B \to \infty$,

$$\widehat{\tau}^2 = \frac{1}{n}\sum_j (T_j - \overline{T})^2 \xrightarrow{P} \tau^2(P_n)$$

where $\overline{T} = \frac{1}{B}\sum_j T_j$. Note that there are two things going on:

1. We estimate $\tau^2(P)$ with $\tau^2(P_n)$
2. We approximate $\tau^2(P_n)$ with the Monte Carlo approximation $\widehat{\tau}^2$.

These are two distinct ideas. The first is plug-in estimation an the second is Monte Carlo.

How do we draw a sample from P_n? Remember that P_n puts mass $1/n$ at eachy data point. The distribution looks like this:

value	X_1	X_2	$\cdots$	X_n
mass	$1/n$	$1/n$	$\cdots$	X_n

To draw X_1^* we just draw one datapoint at random. To draw X_2^* we again draw one datapoint at random. We repeat this n times to get one bootstrap sample. Note that this is equivalent to drawing n times from the data *with replacement*. Draw a point; put it back; drsw a point; put it back; etc. For this reason, people often describe drawing a bootstrap sample as *resampling the data*. But is is best regarded as drawing n times from P_n.

Now we can use the bootstrap for statistical inference. Suppose that $\widehat{\psi}_n = g(X_1, \ldots, X_n)$ is an estimator. For example, it could be a plug-in estimator. Now we apply the bootstrap method. We sample n observations from P_n and re-compute the estimator. We repeat B times to get $\widehat{\tau}$ which is the estimated standard error of $\widehat{\psi}_n$.

3 Bootstrap Confidence Intervals

We can also use the bootstrap to get a confidence interval for ψ. In fact, I will describe three methods.

Method 1. If $\widehat{\psi}_n$ is asymptotically Normal then a $1-\alpha$ confidence interval is

$$\widehat{\psi}_n \pm z_{\alpha/2}\widehat{\tau}$$

where $\widehat{\tau}$ is the bootstrap estimate of the standard error.

Method 2: The Percentile Interval. Let $\widehat{\psi}_1^*, \ldots, \widehat{\psi}_B^*$ denote the bootstrap values of the estimator. The percentile confidence interval is

$$C_n = [\widehat{\psi}^*_{(\alpha/2)}, \widehat{\psi}^*_{(1-\alpha/2)}]$$

where $\widehat{\psi}^*_{(\alpha/2)}$ is the $\alpha/2$ quantile of $\widehat{\psi}^*_1, \ldots, \widehat{\psi}^*_B$ and $\widehat{\psi}^*_{(1-\alpha/2)}$ is the $1-\alpha/2$ quantile of $\widehat{\psi}^*_1, \ldots, \widehat{\psi}^*_B$.

Method 3: The Basic Bootstrap (Reverse Perentile). Suppose for a moment that we knew the distribution

$$G_n(t) = P(\sqrt{n}(\widehat{\psi}_n - \psi) \leq t).$$

Let $g_{\alpha/2} = G_n^{-1}(\alpha/2)$ and $g_{1-\alpha/2} = G_n^{-1}(1-\alpha/2)$. Let

$$C_n = \left[\widehat{\psi}_n - \frac{g_{1-\alpha/2}}{\sqrt{n}}, \widehat{\psi}_n - \frac{g_{\alpha/2}}{\sqrt{n}}\right].$$

Now

$$\begin{aligned}\mathbb{P}(\psi \in C_n) &= \mathbb{P}\left(g_{\alpha/2} \leq \sqrt{n}(\widehat{\psi}_n - \psi) \leq g_{1-\alpha/2}\right) \\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha.\end{aligned}$$

This interval looks strange because you are used to Normal-based intervals. In fact, if G_n is Normal, this interval can be re-written to look like the usual interval due to the symmetry of the Normal.

We do not know G_n so we can't use this interval. But we can estimate G_n with the bootstrap. We define

$$\widehat{G}_n(t) = \frac{1}{n}\sum_{j=1}^{B} I(\sqrt{n}(\widehat{\psi}^*_j - \widehat{\psi}) \leq t).$$

We then estimate $g_{\alpha/2} = G_n^{-1}(\alpha/2)$ and $g_{1-\alpha/2} = G_n^{-1}(1-\alpha/2)$ with $\widehat{g}_{\alpha/2} = \widehat{G}_n^{-1}(\alpha/2)$ and $\widehat{g}_{1-\alpha/2} = \widehat{G}_n^{-1}(1-\alpha/2)$. The confidence interval is

$$C_n = \left[\widehat{\psi}_n - \frac{\widehat{g}_{1-\alpha/2}}{\sqrt{n}}, \widehat{\psi}_n - \frac{\widehat{g}_{\alpha/2}}{\sqrt{n}}\right].$$

Note that

$$\widehat{g}_{(\alpha/2)} = \sqrt{n}(\widehat{\psi}^*_{(\alpha/2)} - \widehat{\psi})$$

and

$$\widehat{g}_{1-(\alpha/2)} = \sqrt{n}(\widehat{\psi}^*_{1-(\alpha/2)} - \widehat{\psi})$$

so that

$$\widehat{\psi} - \frac{\widehat{g}_{1-(\alpha/2)}}{\sqrt{n}} = 2\widehat{\psi} - \widehat{\psi}^*_{1-(\alpha/2)}$$

and

$$\widehat{\psi} - \frac{\widehat{g}_{(\alpha/2)}}{\sqrt{n}} = 2\widehat{\psi} - \widehat{\psi}^*_{(\alpha/2)}.$$

Therefore, we can write

$$C_n = \left[2\widehat{\psi} - \widehat{\psi}^*_{1-(\alpha/2)}, 2\widehat{\psi} - \widehat{\psi}^*_{(\alpha/2)}\right].$$

Again, it looks weird but it follows from the calculations.

4 The Parametric Bootstrap

The bootstrap can also be used for parametric models. Instead of drawing $X_1^*, \ldots, X_n^* \sim P_n$ we instead draw $X_1^*, \ldots, X_n^* \sim p(x; \widehat{\theta})$. The res is the same.

5 Variants

There are many many many papers that have been written about the bootstrap. There are many different versions: the block bootstrap for time-series, the residual bootstrap or the wild bootstrap for regression, the smooth bootstrap, the bias-corrected bootstrap, and many others.

6 Why Does the Bootstrap Work?

We want that the quantiles of the bootstrap distribution of our statistic should be close to the quantiles its actual distribution. Let

$$\widehat{F}_n(t) = \mathbb{P}_n(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \leq t | X_1, \ldots, X_n),$$

be the CDF of the bootstrap distribution, and

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t),$$

be the CDF of the true sampling distribution of our statistic. We want to show that

$$\sup_t |\widehat{F}_n(t) - F_n(t)| \to 0.$$

This turns out to be true in quite a bit of generality, only requiring mild conditions (Hadamard differentiability) but we will prove it in the simplest case: when $\widehat{\theta}_n$ is a sample mean. In this case there are much simpler ways to construct confidence intervals (using Normal approximations) but that is not really the point.

Suppose that $X_1, \ldots, X_n \sim P$ where X_i has mean μ and variance σ^2. Suppose we want to construct a confidence interval for μ.

Let $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and define

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t). \tag{1}$$

We want to show that

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \leq t \;\Big|\; X_1, \ldots, X_n\Big)$$

is close to F_n.

Theorem 1 (Bootstrap Theorem) *Suppose that* $\mu_3 = \mathbb{E}|X_i|^3 < \infty$. *Then,*

$$\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

To prove this result, let us recall that Berry-Esseen Theorem.

Theorem 2 (Berry-Esseen Theorem) *Let* $X_1, \ldots, X_n$ *be i.i.d. with mean* μ *and variance* σ^2. *Let* $\mu_3 = \mathbb{E}[|X_i - \mu|^3] < \infty$. *Let* $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ *be the sample mean and let* Φ *be the cdf of a* $N(0,1)$ *random variable. Let* $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. *Then*

$$\sup_z \Big|\mathbb{P}(Z_n \leq z) - \Phi(z)\Big| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}. \tag{2}$$

Proof of the Bootstrap Theorem. Let $\Phi_\sigma(t)$ denote the cdf of a Normal with mean 0 and variance σ^2. Let $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$. Thus, $\widehat{\sigma}^2 = \mathrm{Var}(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n)|X_1, \ldots, X_n)$. Now, by the triangle inequality,

$$\begin{aligned}\sup_t |\widehat{F}_n(t) - F_n(t)| &\leq \sup_t |F_n(t) - \Phi_\sigma(t)| + \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| + \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

Let $Z \sim N(0,1)$. Then, $\sigma Z \sim N(0, \sigma^2)$ and from the Berry-Esseen theorem,

$$\begin{aligned}\mathrm{I} &= \sup_t |F_n(t) - \Phi_\sigma(t)| = \sup_t \Big|\mathbb{P}\left(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t\right) - \mathbb{P}\left(\sigma Z \leq t\right)\Big| \\ &= \sup_t \left|\mathbb{P}\left(\frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{\sigma} \leq \frac{t}{\sigma}\right) - \mathbb{P}\left(Z \leq \frac{t}{\sigma}\right)\right| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}.\end{aligned}$$

Using the same argument on the third term, we have that

$$\mathrm{III} = \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \leq \frac{33}{4}\frac{\widehat{\mu}_3}{\widehat{\sigma}^3\sqrt{n}}$$

where $\widehat{\mu}_3 = \frac{1}{n}\sum_{i=1} |X_i - \widehat{\mu}_n|^3$ is the empirical third moment. By the strong law of large numbers, $\widehat{\mu}_3$ converges almost surely to μ_3 and $\widehat{\sigma}$ converges almost surely to σ. So, almost surely, for all large n, $\widehat{\mu}_3 \leq 2\mu_3$ and $\widehat{\sigma} \geq (1/2)\sigma$ and $\mathrm{III} \leq \frac{33}{4}\frac{4\mu_3}{\sqrt{n}}$. From the fact that $\widehat{\sigma} - \sigma = O_P(\sqrt{1/n})$ it may be shown that $\mathrm{II} = \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| = O_P(\sqrt{1/n})$. (This may be seen by Taylor expanding $\Phi_{\widehat{\sigma}}(t)$ around σ.) This completes the proof. □

So far we have focused on the mean. Similar theorems may be proved for more general parameters. The details are complex so we will not discuss them here.

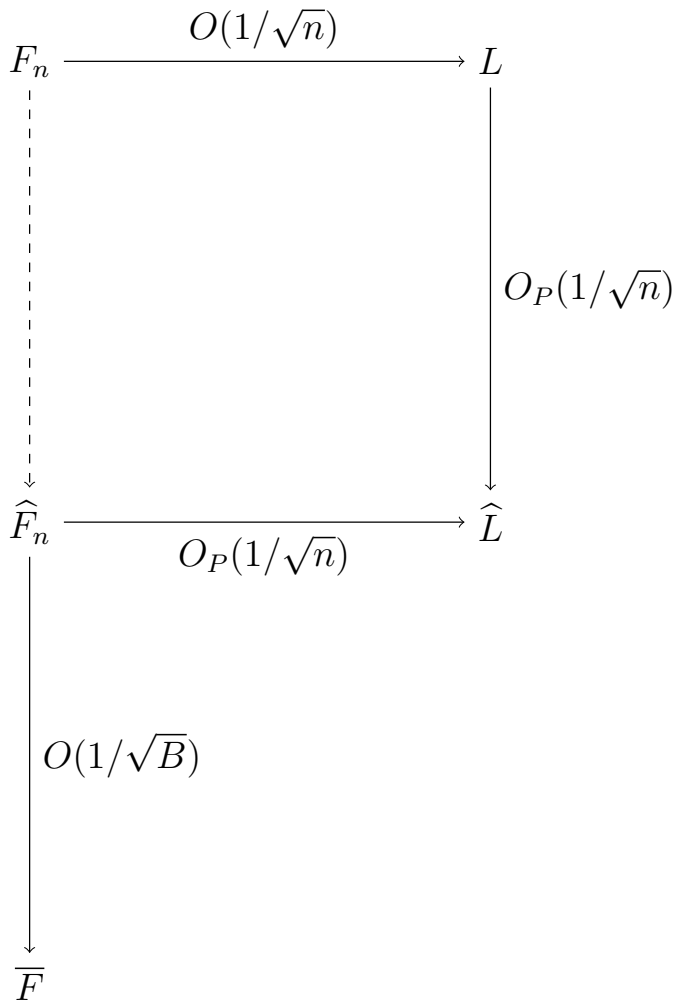

Figure 1: *The distribution* $F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t)$ *is close to some limit distribution* L*. Similarly, the bootstrap distribution* $\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}^*_n - \widehat{\theta}_n) \leq t | X_1, \ldots, X_n)$ *is close to some limit distribution* $\widehat{L}$*. Since* $\widehat{L}$ *and* L *are close, it follows that* F_n *and* $\widehat{F}_n$ *are close. In practice, we approximate* $\widehat{F}_n$ *with its Monte Carlo version* $\overline{F}$ *which we can make as close to* $\widehat{F}_n$ *as we like by taking* B *large.*

7 Failure of the Bootstrap

As usual when we need a counterexample we try the uniform distribution. Suppose that $X_1, \ldots, X_n \sim U[0, \theta]$ and we try to bootstrap the MLE to construct a confidence interval for θ. The mle is $X_{(n)}$. This point is contained in the bootstrap sample with probability

$$1 - (1 - 1/n)^n \approx .63.$$

So the bootstrap distribution puts mass .63 at the single point $X_{(n)}$. But we know that $n(X_{(n)} - \theta)$ has an exponential distribution. So the bootstrap distribution does not resemble the true distribution.