36-705: Intermediate Statistics **Fall 2019**

Lecture 26: November 4

Lecturer: Siva Balakrishnan

We continue our discussion of confidence intervals and then turn our attention to the bootstrap.

26.1 Inverting Probability Inequalities

In some simple cases, we can use tail bounds to derive confidence intervals. These typically have the advantage of being exact, finite-sample intervals. However, they are rarely used in practice for many reasons including: (1) we do not always have tail bounds for estimators of interest (2) there are usually imprecisely known constants in tails bounds (3) related to (2) they are often very conservative (i.e. the intervals are often too wide to be useful).

Here are a couple of examples:

Example 26.1 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. By Hoeffding's inequality:*

$$\mathbb{P}(|\widehat{p} - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

Then

$$\mathbb{P}\left(|\widehat{p} - p| > \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence, $\mathbb{P}(p \in C) \geq 1 - \alpha$ where $C = (\widehat{p} - \epsilon_n, \widehat{p} + \epsilon_n)$.

Example 26.2 *Let $X_1, \ldots, X_n \sim F$. Suppose we want a* **confidence band** *for F. We can use VC theory. Remember that*

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

Then

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence,

$$P_F(L(t) \leq F(t) \leq U(t) \text{ for all } t) \geq 1 - \alpha$$

for all F, where

$$L(t) = \widehat{F}_n(t) - \epsilon_n, \quad U(t) = \widehat{F}_n(t) + \epsilon_n.$$

We can improve this by taking

$$L(t) = \max\left\{\widehat{F}_n(t) - \epsilon_n,\ 0\right\}, \quad U(t) = \min\left\{\widehat{F}_n(t) + \epsilon_n,\ 1\right\}.$$

26.1.1 Pivots

Another useful way of attempting to construct confidence intervals is to base the intervals on *pivots*. A pivot is a function of the data and the unknown parameter θ – $Q(X_1, \ldots, X_n, \theta)$ – whose distribution does not depend on θ.

Let us consider two examples:

1. Suppose that $X_1, \ldots, X_n \sim N(\theta, 1)$ then we can see that $Q(X_1, \ldots, X_n) = \overline{X_n} - \theta \sim N(0, 1/n)$ and so the distribution of Q does not depend on θ.
2. Suppose we consider $X_1, \ldots, X_n \sim U[0, \theta]$ and we consider the function:

   $$Q(X_1, \ldots, X_n, \theta) = \frac{\max_i X_i}{\theta},$$

   has distribution:

   $$P(Q(X_1, \ldots, X_n, \theta) \leq t) = \begin{cases} t^n & 0 \leq t \leq 1 \\ 1 & t \geq 1. \end{cases}$$

   Once again the distribution does not depend on θ.

Given a pivot we can construct confidence intervals in a simple way. Since the distribution of Q does not depend on θ, we can find a, b which do not depend on θ such that:

$$\mathbb{P}_\theta(a \leq Q(X_1, \ldots, X_n, \theta) \leq b) = 1 - \alpha, \quad \text{for all } \theta \in \Theta.$$

Now, we construct our confidence interval as:

$$C(X_1, \ldots, X_n) = \{\theta : a \leq Q(X_1, \ldots, X_n, \theta) \leq b\}.$$

By our construction:

$$\mathbb{P}_\theta(\theta \in C(X_1, \ldots, X_n)) = \mathbb{P}_\theta(a \leq Q(X_1, \ldots, X_n, \theta) \leq b) = 1 - \alpha.$$

Going back to our two examples we find that we will once again obtain the now standard intervals for the two problems (the additive interval for the Gaussian mean, and the multiplicative scale interval for the uniform parameter).

26.2 Tests Versus Confidence Intervals

Confidence intervals are more informative than tests. Intuitively, p-values are more informative than an accept/reject decision because it summarizes all the significance levels for which we would reject the null hypothesis. Similarly, a confidence interval is more informative that a test because it summarizes all the parameters for which we would (fail to) reject the null hypothesis. More practically, a confidence interval tells us something about the "effect size" as well as something about the uncertainty in our estimate of the "effect size".

Look at Figure 25.1. Suppose we are testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. We see 5 different confidence intervals. The first two cases (top two) correspond to not rejecting H_0. The other three correspond to rejecting H_0. Reporting the confidence intervals is much more informative than simply reporting "reject" or "don't reject."

26.3 Bootstrap samples

We have discussed this before when we discussed plug-in estimators: given samples $X_1, \ldots, X_n \sim P$ we can write the empirical CDF as:

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x),$$

and the corresponding empirical distribution as:

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \in A).$$

We can also imagine drawing *bootstrap samples* by drawing samples from P_n. We denote these as:

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

Drawing from the empirical distribution is the same as drawing from the distribution that puts mass $1/n$ at each observed sample, i.e. it is the same as drawing from the uniform distribution on the given samples. Equivalently, you can imagine drawing from the given samples (uniformly) with replacement.

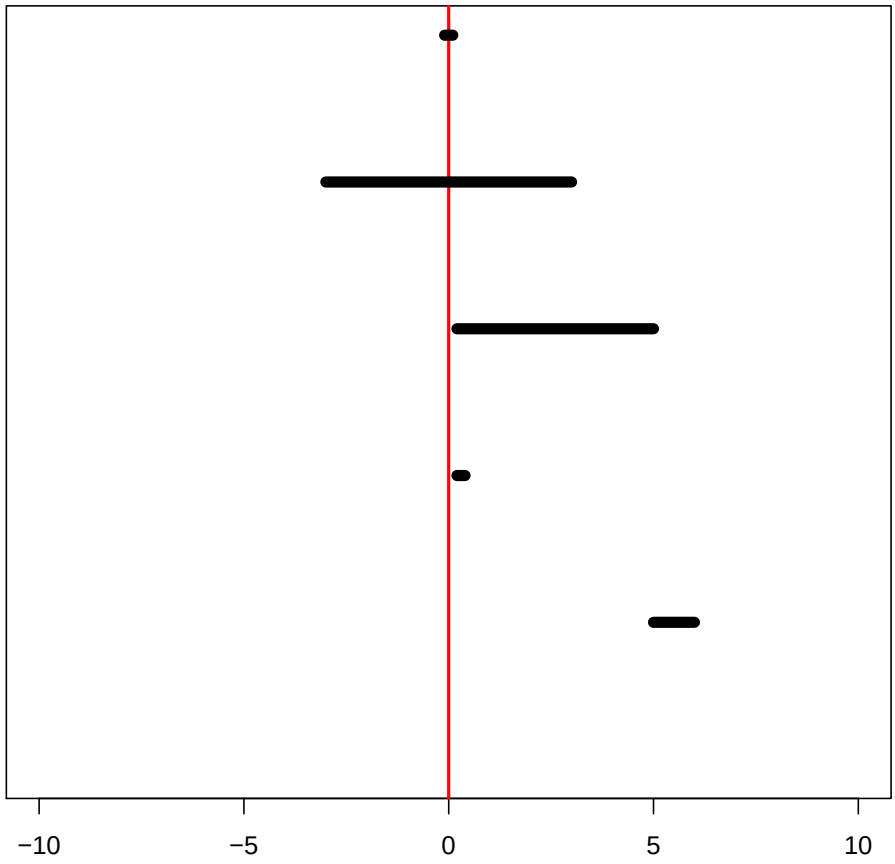

Figure 26.1: Five examples: 1. Not significant, precise. 2. Not significant, imprecise. 3. Barely significant, imprecise. 4. Barely significant, precise. 5. Significant and precise.

26.4 Bootstrap variance estimate

To understand the idea, let us first consider the Monte-Carlo variance estimate. Suppose we had an estimator $\widehat{\theta}_n = g(X_1, \ldots, X_n)$ (this could be a complicated function), where $X_1, \ldots, X_n \sim P$ and we want to estimate $\mathrm{Var}_P(\widehat{\theta}_n)$.

Supposing that we knew P we could try to compute the variance analytically: this might be difficult. The Monte-Carlo variance estimate would be to instead draw B samples of size n from P, i.e. we draw, $\{X_{11}, \ldots, X_{1n}\}, \ldots, \{X_{B1}, \ldots, X_{Bn}\} \sim P$, to compute our estimator on each of these samples, i.e. compute $\widehat{\theta}_n^{(1)}, \ldots, \widehat{\theta}_n^{(B)}$ and then use the sample variance, i.e.

$$\widehat{\sigma}_n^2 = \frac{1}{B}\sum_{i=1}^{B}\left(\widehat{\theta}_n^{(i)}\right)^2 - \left(\frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_n^{(i)}\right)^2.$$

By the LLN we have that $\widehat{\sigma}^2 \xrightarrow{p} \mathrm{Var}_P(\widehat{\theta}_n)$. Unfortunately, we typically do not know P.

By now, you have already guessed the idea behind the bootstrap. The idea is to replace P in the above procedure by the empirical distribution P_n. We'll reason about this more carefully in the next lecture. For now, here is the algorithm:

Bootstrap Variance Estimator

1. Draw a bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\theta}_n^* = g(X_1^*, \ldots, X_n^*)$.
2. Repeat the previous step, B times, yielding estimators $\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,B}^*$.
3. Compute:
$$\widehat{s}^2 = \frac{1}{B}\sum_{j=1}^{B}(\widehat{\theta}_{n,j}^* - \overline{\theta})^2,$$
where $\overline{\theta} = \frac{1}{B}\sum_{j=1}^{B}\widehat{\theta}_{n,j}^*$.
4. Output $\widehat{s}^2$.

26.5 Bootstrap Confidence Intervals

The bootstrap can also be used to obtain confidence intervals. If your estimator has a normal limit then you could just use a Wald interval with the bootstrap variance estimate, i.e. $C_n = [\widehat{\theta}_n - \widehat{s}z_{\alpha/2}, \widehat{\theta}_n + \widehat{s}z_{\alpha/2}]$.

It is often more accurate to use the distribution of the bootstrap estimates itself to construct the bootstrap confidence interval.

26.5.1 Hypothetical confidence interval

Suppose we knew the distribution of our estimator, in particular suppose we knew the distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$. Let us denote the distribution by G and denote its $\alpha/2$ and $1-\alpha/2$ quantiles by $g_{\alpha/2}$ and $g_{1-\alpha/2}$.

Then a $1-\alpha$ confidence interval would be:

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

This might seem a little strange, but this is probably because you are used to confidence intervals based on the normal distribution which has symmetric quantiles. To verify this,

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

Again the point is that we do not know the distribution G above so we try to approximate this using the bootstrap.

26.5.2 Bootstrap confidence interval algorithm

Bootstrap Confidence Interval

1. Draw a bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\theta}_n^* = g(X_1^*, \ldots, X_n^*)$.
2. Repeat the previous step, B times, yielding estimators $\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,B}^*$.
3. Let
$$\widehat{G}(t) = \frac{1}{B}\sum_{j=1}^{B} I\left(\sqrt{n}(\widehat{\theta}_{n,j}^* - \widehat{\theta}_n\right) \leq t).$$
4. Let
$$C_n = \left[\widehat{\theta}_n - \frac{g_{1-\alpha/2}}{\sqrt{n}},\ \widehat{\theta}_n - \frac{g_{\alpha/2}}{\sqrt{n}}\right]$$
where $g_{\alpha/2} = \widehat{G}^{-1}(\alpha/2)$ and $g_{1-\alpha/2} = \widehat{G}^{-1}(1-\alpha/2)$.
5. Output C_n.

26.6 Variants

There are many many many papers that have been written about the bootstrap. Particularly, there are lots of variants – the block bootstrap for time-series, the residual bootstrap or the wild bootstrap for regression, the parametric bootstrap for parametric models, the smooth bootstrap and ideas related to sub-sampling to avoid certain regularity conditions, the less computationally intensive but less general Jackknife and so on.