6.S951 Modern Mathematical Statistics Fall 2024

Lecture 11 — October 17, 2024

Prof. Stephen Bates

Scribe: Oliver Wang

1 Outline

Agenda: Bootstrap

1. Empirical Distribution, Plug-in idea
2. Bootstrap variance estimation
3. Bootstrap "pivot" CIs

Last time:

1. Inference for means (CLT, Hoeffding)
2. Inference (confidence bands) for CDFs (DKW, simulation)

2 Empirical Distributions

The core idea here is that given samples/observations $X_i \in \mathcal{X} \overset{\text{iid}}{\sim} P,\ i = 1, ..., n$, one can view $\vec{X} = (X_1, ..., X_n)$ as a (empirical) distribution.

Definition 1 (Empirical Distribution)**.** *The empirical distribution is* $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ *where* δ_{x_i} *is a point mass at* x_i*.*

Note that if $\mathcal{X} = \mathbb{R}$ the empirical CDF $\hat{F}$ corresponds to $\hat{P}$.

Through the definition of observed samples as a distribution, we can then approximate the true distribution P with $\hat{P}$ while $P \neq \hat{P}$. The perspective lend itself naturally to asymptotics for which we can denote a n-dependent $\hat{P}_n$.

Moving on, let us define statistics for functions

Definition 2 (Statistical Functional)**.** *A statistical functional is a mapping*

$$T : \{\text{distribution on } \mathcal{X}\} \to \mathbb{R}^d. \tag{1}$$

The point here is this mapping also applies to infinite data points as a distribution

Example 2.1.

$$\text{Mean} : T(P) = \int x dP_x$$
$$\text{Sample mean } : T(\hat{P}) = \overline{X}$$

Example 2.2.

$$\text{Variance} : T(P) = \int x^2 dP_x - \Big(\int x dP_x\Big)^2$$
$$\text{Empirical variance } : T(\hat{P}) = \overline{X^2} - (\overline{X})^2$$

Example 2.3.

$$\text{Least-squares} : T(P) = \arg\min_{\theta} \ \mathbb{E}_P(Y - \theta^\top X)^2$$
$$\text{Least-squares estimator } : T(\hat{P}) = \arg\min_{\theta} \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta^\top X_i)^2$$

In general, $T(\hat{P})$ is a reasonable approximate to $T(P)$ since $\hat{P} \approx P$. This is known as the "Plug-in" principle, which will lead to Bootstrap.

3 Bootstrap Variance Estimation

Given observation: $\vec{X} = (X_1, ..., X_n)$, we want to estimate $T(P) \in \mathbb{R}^d$.

Moreover, suppose that $T' : \mathcal{X}^n \to R^d$ is an estimation, we wish to understand the variance $\text{Var}_P(T'(\vec{X}))$. One can think of such T' as $T(\hat{P})$ but the idea generalizes.

For intuition, the observations are randomly given by underlying universe using a random seed, want to understand how much do these sample vary under varying seeds.

The Bootstrap idea: Given the empirical distribution as samples, we posit that

$$\text{Var}_P(T'(\vec{X})) \approx \text{Var}_{\hat{P}}(T'(\vec{X})) \tag{2}$$

where $\text{Var}_{\hat{P}}(T'(X))$ can be computed via simulations like Monte Carlo. Useful when the analytical computation of the RHS is difficult.

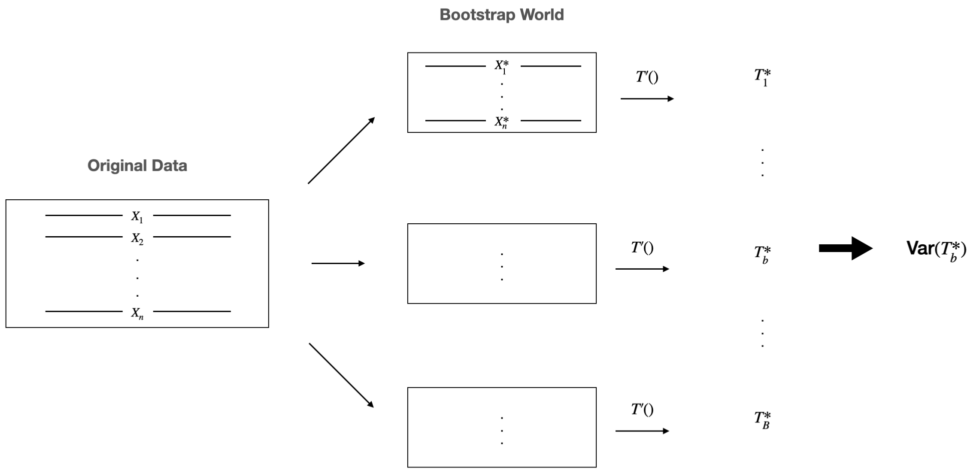
A quick notation change: $\widehat{\text{Var}}(T') : \mathcal{X}^n \to \mathbb{R}^{d\times d} := \text{Var}_{\hat{P}}(T'(\vec{X}))$.

Now, we define a algorithm for computing $\widehat{\text{Var}}$:

Algorithm 1 Bootstrap for variance estimation

for b $= 1, ..., B$ **do**
  draw $X_i^* \overset{\text{iid}}{\sim} \hat{P}$, i=1,...,n
  compute $T_b^* = T'(X_1^*, ..., X_n^*)$
end for
return $\hat{\text{Var}} = \text{Var}(T_1^*, ..., T_B^*) = \frac{1}{B}\sum_b(T_b^* - \overline{T^*})$

The extra randomness for variance estimation here is really from the Monte Carlo simulation. To aid understanding, here is illustration for the algorithm



We can also construct CIs for $\hat{\text{Var}}$:

$$C(\vec{X}) = T'(\vec{X}) \pm \Phi^{-1}\big(1 - \frac{\alpha}{2}\big)\widehat{\text{Var}}(T')$$

assuming that T' is centered correctly and the variance estimation is reasonably well. An associated fact: $P(T(P) \in C(\vec{X})) \xrightarrow{P} 1 - \alpha$ as $n \to \infty$ under general conditions.

Note that Bootstrapping works even for misspecified models since the workflow does not require model correctness. One can also perform Sub-sampling when the target variance estimation is not the infinite dimensional truth but a finite sample empirical.

4 Pivotal Bootstrap

From the slides, when the sampling distribution is biased or skewed or both, we hope to adjust the CIs to account for these discrepancies. The idea for this is we will use Bootstrap to approximate the distribution of $T'(\vec{X}) - T(P) \sim Q$. We provide an algorithm for this:

Algorithm 2 Bootstrap approximation to Q

for b $= 1, ..., B$ **do**
  draw $X_i^* \overset{\text{iid}}{\sim} \hat{P}$
  compute $R_b^* = T(X_1^*, ..., X_n^*) - T(\hat{P})$
end for
return $\hat{Q} = \frac{1}{B}\sum_b \delta_{R_b^*}$

We can then construct CI using the quantile of $\hat{Q}$.

Pivotal CIs : If $T'(X) - T(P) \dot{\sim} Q$, then

$$P\Big(q_{\frac{\alpha}{2}} \leq T'(\vec{X}) - T(P) \leq q_{1-\frac{\alpha}{2}}\Big) \approx 1 - \alpha. \tag{3}$$

where q denote the quantile for the empirical distribution $\hat{Q}$. We can then obtain the pivotal CI

$$T'(\vec{X}) - q_{1-\frac{\alpha}{2}} \leq T(P) \leq T'(\vec{X}) - q_{\frac{\alpha}{2}}$$

which is adjusted for bias and skew. The idea is that we want to construct CIs using distributions of some already-stable statistics.