Outline

1) Nonparametric Estimation
2) Plugin estimator
3) Bootstrap standard errors
4) Bootstrap bias estimator / correction
5) Bootstrap confidence intervals
6) Double bootstrap
Setting Nonparametric iid sampling model  
\( X_1, \ldots, X_n \sim P, \ P \text{ unknown} \)

Want to do inference on some "parameter" \( \Theta(P) \)

\[
\begin{align*}
\text{Ex a) } & \Theta(P) = \text{median}(P) \quad (X \in \mathbb{R}) \\
\text{b) } & \Theta(P) = \lambda_{\max}(\text{Var}_P(x_i)) \quad (X \in \mathbb{R}^d) \\
\text{c) } & \Theta(P) = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_P \left[ (Y_i - \theta'x_i)^2 \right] \\
\text{d) } & \Theta(P) = \arg\min_{\theta \in \Theta} D_{KL}(P \parallel P_0) \quad \text{(best-fitting model even if misspec)} \equal\ \arg\max_{\theta} \mathbb{E}_P \left[ \ell_i(\theta; x_i) \right]
\end{align*}
\]

Recall the empirical dist. of \( X_1, \ldots, X_n \) is  
\( \hat{P}_n = \frac{1}{n} \sum \delta_{X_i} \quad (\hat{P}_n(A) = \frac{\#\{i : X_i \in A\}}{n}) \)

The plug-in estimator of \( \Theta(P) \) is \( \hat{\theta} = \Theta(\hat{P}_n) \)

a) Sample median  
b) \( \lambda_{\max}(\text{sample var}) \)  
c) OLS estimator  
d) MLE for \( \exists P_0 : \theta \in \Theta \)
Does plug-in estimator work? Depends
\( \hat{P}_n \overset{p}{\to} P ? \) Dep. on what sense of convergence

\[ \hat{P}_n(A) \overset{p}{\to} P(A) \text{ for all } A \]

\( \sup_A |\hat{P}_n(A) - P(A)| \overset{P}{\to} 0 \) if \( P \text{ cts } X \)

(use \( A_n = \{X_1, \ldots, X_n\} \))

\[ \sup_x |\hat{P}_n((-\infty, x]) - P((-\infty, x])| \overset{P}{\to} 0 \text{ for } x \in \mathbb{R} \]

Want \( \Theta(P) \) to be cts wrt some topology
in which \( \hat{P}_n \overset{P}{\to} P \), then \( \Theta(\hat{P}_n) \overset{P}{\to} \Theta(P) \)

Counterexamples

\[ \Theta(P) = \{ P \text{ is absolutely cts } \} \quad (P \ll \text{Lebesgue}) \]

\[ \Theta(P) = \{ P \text{ is integrable } \} \quad (\int |P(x)| < \infty) \]

\( \hat{P}_n \text{ always integrable, never abs. cts. } \), for all \( n \).
Bootstrap standard errors

Suppose \( \hat{\Theta}_n(X) \) is an estimator for \( \Theta(P) \)

(maybe plug-in, maybe not)

What is its standard error? Use plug-in:

\[
\text{s.e.}(\hat{\Theta}_n) = \sqrt{\text{Var}_{\hat{\pi}_n}(\hat{\Theta}^*_n)} \quad \quad \quad \text{[use } \hat{\Theta}^*_n \text{ to indicate new sample } X^*_n \text{, not } X]\n\]

\[
\text{Var}_{\hat{\pi}_n}(\hat{\Theta}^*_n) = \text{Var}_{X_1^*, \ldots, X_n^*) \sim \hat{\pi}_n}(\hat{\Theta}(X_1^*, \ldots, X_n^*))
\]

How to compute? Monte Carlo:

For \( b = 1, \ldots, B \):

\begin{align*}
\text{Sample } X_1^b, \ldots, X_n^b \sim \hat{\pi}_n \\
\hat{\Theta}^b = \hat{\Theta}(X_1^b, \ldots, X_n^b) \\
\overline{\Theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\Theta}^b \\
\text{s.e.}(\hat{\Theta}_n) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (\hat{\Theta}^b - \overline{\Theta}^*)^2}
\end{align*}

Note this is a Monte Carlo numerical approx. to the idealized Bootstrap estimator, which we could compute by iterating over all \( n^n \) possible \( X^* = (X_1^*, \ldots, X_n^*) \) vectors.
Bootstrap Bias Correction

\( \hat{\theta}_n \) some estimator. What is its bias?

\[
\text{Bias}_{\text{p}}(\hat{\theta}_n) = \mathbb{E}_{\text{p}} \left[ \hat{\theta}_n - \theta(\text{p}) \right]
\]

Idea: plug in \( \hat{\theta}_n \) for \( \text{p} \):

\[
\text{Bias}_{\hat{\theta}_n}(\hat{\theta}_n^*) = \mathbb{E}_{\hat{\theta}_n} \left[ \hat{\theta}_n^* - \theta(\hat{\theta}_n) \right]
\]

Monte Carlo:

For \( b = 1, \ldots, B \):

Sample \( X_1^b, \ldots, X_n^b \overset{iid}{\sim} \hat{\theta}_n \)

\( \hat{\theta}_n^b = \hat{\theta}(X_n^b) \)

\( \bar{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_n^b \)

\( \hat{\text{Bias}}(\hat{\theta}_n) = \bar{\theta}^* - \theta(\hat{\theta}_n) \)

We can use this to correct bias:

\( \hat{\theta}_n^{\text{bc}} = \hat{\theta}_n - \hat{\text{Bias}}(\hat{\theta}_n) \)

Note: while \( \hat{\theta}_n - \hat{\text{Bias}}(\hat{\theta}_n) \) is always better than \( \hat{\theta}_n \), \( \hat{\theta}_n - \hat{\text{Bias}}(\hat{\theta}_n) \) may not be! Might be adding uncorrelated noise.
**Sampling dist.**

Parameter

Data set

Estimator

Sampling dist. of estimator

---

**Real World**

\( P = \bigwedge \Theta(P) \)

\( X_1, \ldots, X_n \stackrel{iid}{\sim} P \) (observed once)

\( \hat{\Theta}(X) \)

---

**Bootstrap World**

\( \hat{P}_n(x) = \)  

\( \Theta(\hat{P}_n(x)) \)

\( X^*_1, \ldots, X^*_n \stackrel{iid}{\sim} \hat{P}_n(x) \)

\( \hat{\Theta}^* = \hat{\Theta}(X^*) \) (generated at will)

---

**Parameter**

Bias

Bias

Bias

Bias

Bias
Bootstrap Confidence Interval

How do we get a CI for \( \Theta(P) \)?

Idea: What if we knew the distribution of \( R_n(x, P) = \hat{\Theta}_n(x) - \Theta(P) \)?

Define cdf \( G_{n,P}(r) = P(\hat{\Theta}(x) - \Theta(P) \leq r) \)

Lower \( \frac{\alpha}{2} \) quantile \( r_1 = G_{n,P}^{-1}(\frac{\alpha}{2}) \)

Upper \( \frac{\alpha}{2} \) quantile \( r_2 = G_{n,P}^{-1}(1 - \frac{\alpha}{2}) \)

\[ 1 - \alpha = P_P(\hat{\Theta}_n - \Theta \leq r_2) \]

\[ = P_P(\Theta \in [\hat{\Theta}_n - r_2, \hat{\Theta}_n - r_1]) \]

Usually we don't know \( G_{n,P} \) -- so bootstrap!

\( G_{n,\hat{P}_n}(r) = P_{\hat{P}_n}(\hat{\Theta}(x^*) - \Theta(\hat{P}_n) \leq r) \)

\( G_{n,\hat{P}_n}(r) \) is a function only of \( X \) (not of \( P \))

Can use \( C_{n,\alpha} = [\hat{\Theta}_n - \hat{r}_2, \hat{\Theta}_n - \hat{r}_1] \)

with \( \hat{r}_1 = G_{n,\hat{P}_n}^{-1}(\frac{\alpha}{2}) \), \( \hat{r}_2 = G_{n,\hat{P}_n}(1 - \frac{\alpha}{2}) \)
Bootstrap algo:

For \( b=1, \ldots, B \):

\[
X_{i.b}^*, \ldots, X_{n.b}^* \sim \bar{P}_n
\]

\[
R_{n.b}^* = \hat{\Theta}(X_{n.b}) - \Theta(\bar{P}_n)
\]

Return ecdf of \( R_{n.b}^* \)

The quantity \( R_n(X, \mathcal{P}) = \hat{\Theta}(X) - \Theta(\mathcal{P}) \) is called a root (function of data + dist., used to make C.I.s)

Other examples:

\[
R_n(X, \mathcal{P}) = \frac{\hat{\Theta}_n(X) - \Theta(\mathcal{P})}{\hat{\sigma}(X)}
\]

[where \( \hat{\sigma}(X) \) is some estimate of s.e.(\( \hat{\Theta}_n \))]

\[
R_n(X, \mathcal{P}) = \frac{\hat{\Theta}_n(X)}{\Theta(\mathcal{P})}
\]

Want to choose \( R_n \) so its sampling dist.

\( G_{n, \mathcal{P}} \) changes slowly with \( \mathcal{P} \) (so \( G_{n, \hat{\Theta}_n} \approx G_{n, \mathcal{P}} \))

Studentized root \( \frac{\hat{\Theta}_n - \Theta}{\hat{\sigma}} \) usually works better

than \( \hat{\Theta}_n - \Theta \), then we get

\[
C_{n, \alpha} = \left[ \hat{\Theta}_n - \hat{\sigma}_{\hat{\Theta}_n}, \hat{\Theta}_n + \hat{\sigma}_{\hat{\Theta}_n} \right]
\]
Double Bootstrap

We might have theory that tells us, e.g.,

\[ \sup_{a < b} \left| G_{n, \hat{\alpha}}(\left[a, b\right]) - G_{n, \alpha}(\left[a, b\right]) \right|^p \rightarrow 0 \]

but still be worried about finite-sample coverage.

Let \( \gamma_{n, \rho}(\alpha) = \mathbb{P}_\rho \left( C_{n, \alpha} \in \Theta(\mathbb{P}) \right) \)

\[ \rightarrow 1 - \alpha \quad \text{if } C_{n, \alpha} \text{ has asy. coverage} \]

But in finite samples, might have

\[ \gamma_{n, \rho}(\alpha) < 1 - \alpha \]

e.g., "90% interval" has 87% coverage

\[ \gamma_{n, \rho}(0.1) = 0.87 < 0.9 \]

Solution? \underline{Double Bootstrap}!

1. Estimate \( \gamma_{n, \rho}(\cdot) \) via plug-in \( \gamma_{n, \rho_n}(\cdot) \)

2. Use \( C_{n, \hat{\alpha}}(x) \) where \( \hat{\gamma}(\hat{\alpha}) = 1 - \alpha \)

e.g., estimate "92% interval" has 90% coverage, \( \hat{\alpha} = .08 \)
Step 1 algo.

For \( a = 1, \ldots, A \):

\[
X_1^{*a}, \ldots, X_n^{*a} \sim \text{iid } \hat{P}_n^{*a} \\
\hat{P}_n^{*a} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i^{*a}} \\
\text{For } b = 1, \ldots, B:
\]

\[
X_1^{**a,b}, \ldots, X_n^{**a,b} \sim \text{iid } \hat{P}_n^{*a} \\
R_n^{**a,b} = \left( \hat{\Theta}_n(X^{**a,b}) - \Theta(\hat{P}_n^{*a}) \right) / \hat{\sigma}(X^{**a,b}) \\
\hat{G}_n^{*a} = \text{ecdf}(R_n^{**a,1}, \ldots, R_n^{**a,B})
\]

For \( \alpha \in \text{grid} \):

\[
C_{n,\alpha}^{*a} = [\hat{\Theta}_n^{*a} - \hat{\sigma}^{*a} \cdot \alpha(\hat{G}_n^{*a}), \hat{\Theta}_n^{*a} - \hat{\sigma}^{*a} \cdot \alpha(\hat{G}_n^{*a})]
\]

For \( \alpha \in \text{grid} \):

\[
\hat{\gamma}(\alpha) = \frac{1}{A} \sum_{a=1}^A 1\{C_{n,\alpha}^{*a} \in \Theta(\hat{P}_n)\}
\]

\[
\hat{\alpha} = \hat{\gamma}^{-1}(1-\alpha)
\]