Bootstrapping

- Resampling framework
  → Random forest

- Changed the way many statistical analysis may be carried out

Motivation

Recall 95% C.I of population mean given a sample mean $\bar{x}$ (and sample std dev s)

$$\left(\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \right)$$

→ "we are 95% confident that the true population mean is within this range"

- What if we are interested in a different statistic / estimator?

- What if we can't rely on the assumption of an approximately normal dist?

  ↳ i.e) it's more skewed than thought

MC = Theoretical model
B.S = Data

Bootstrapping

Resampling from the data directly w/ replacement

→ Sample of size n
→ Generate large # B of resamples each of size n with replacement

original sample: $x = \{x_1, \ldots, x_n\} \rightarrow \hat{\theta}_0$

$\{x_{11}^*, \ldots, x_{n1}^*\}_1$
$\{x_{21}, \ldots, x_{n}$

3 measures
- $\widehat{Var}(\,)$
- $\widehat{SE}(\hat{\theta})$
- $\widehat{Bias}(\hat{\theta})$

Note

When we bootstrap a sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ we have to keep the pairs together.

↳ Preserve association between x & y

(Do Q 7.9)