

Bootstrap



Limitations of CV

- An 80/20 split throws out a relatively large amount of data if only have, say, 20 examples.
- Test error is informative, but how accurate is this number? (e.g., 3/5 heads vs. 30/50)
- How do I get confidence intervals on statistics like the median or variance of a distribution?
- Instead of the error for the entire dataset, what if I want to study the error for a *particular example* x ?

The Bootstrap: Developed by Efron in 1979.

Bootstrap: basic idea

Given dataset drawn iid samples with CDF F_Z :

$$\mathcal{D} = \{z_1, \dots, z_n\} \stackrel{i.i.d.}{\sim} F_Z$$

We compute a *statistic* of the data to get: $\hat{\theta} = t(\mathcal{D})$

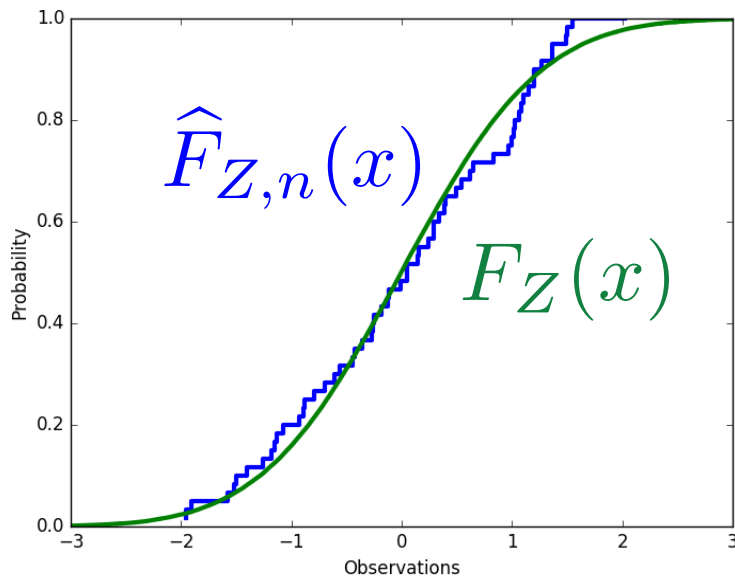
What is the distribution of $\hat{\theta} = t(\mathcal{D})$

Bootstrap: basic idea

Given dataset drawn iid samples with CDF F_Z :

$$\mathcal{D} = \{z_1, \dots, z_n\} \stackrel{i.i.d.}{\sim} F_Z$$

We compute a *statistic* of the data to get: $\hat{\theta} = t(\mathcal{D})$



$$F_Z(x) = \mathbb{P}(Z \leq x)$$

$$\hat{F}_{Z,n}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{z_i \leq x\}$$

$$|\hat{F}_{Z,n}(x) - F_Z(x)| \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.}$$

Bootstrap: basic idea

Given dataset drawn iid samples with CDF F_Z :

$$\mathcal{D} = \{z_1, \dots, z_n\} \stackrel{i.i.d.}{\sim} F_Z$$

We compute a *statistic* of the data to get: $\hat{\theta} = t(\mathcal{D})$

For $b=1, \dots, B$ define the b th **bootstrapped** dataset as drawing n samples **with replacement** from D

$$\mathcal{D}^{*b} = \{z_1^{*b}, \dots, z_n^{*b}\} \stackrel{i.i.d.}{\sim} \hat{F}_{Z,n}$$

and the b th bootstrapped statistic as: $\theta^{*b} = t(\mathcal{D}^{*b})$

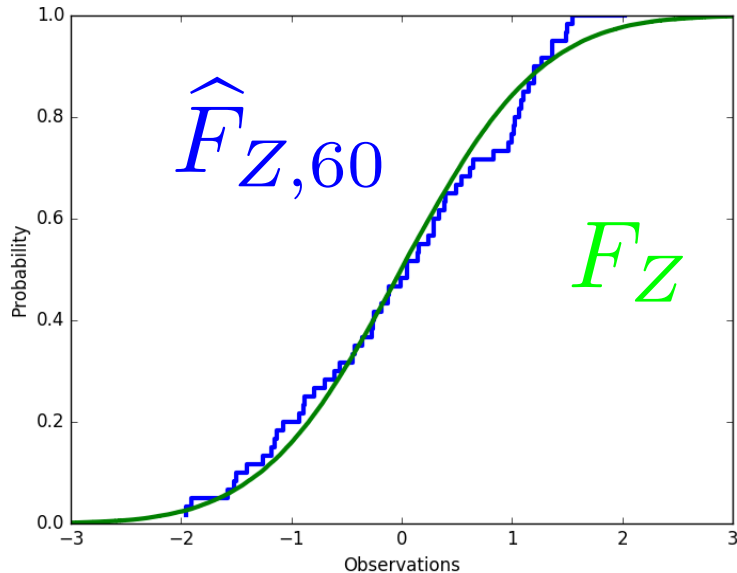
Bootstrap: basic idea

Given dataset drawn iid samples with CDF F_Z :

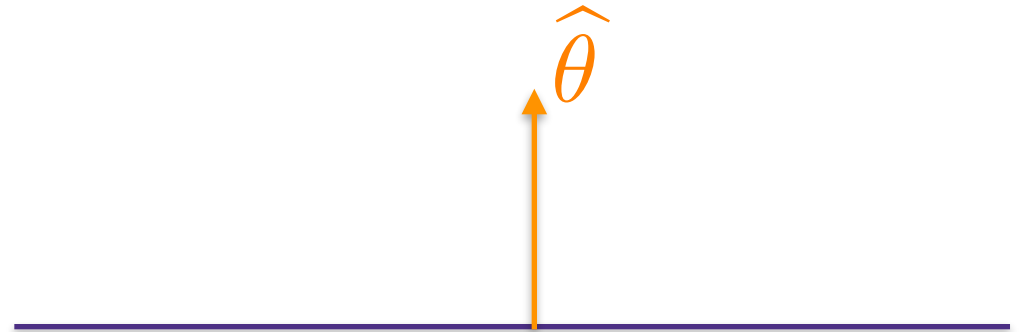
$$\mathcal{D} = \{z_1, \dots, z_n\} \stackrel{i.i.d.}{\sim} F_Z$$

We compute a *statistic* of the data to get: $\hat{\theta} = t(\mathcal{D})$

$$\mathcal{D}^{*b} = \{z_1^{*b}, \dots, z_n^{*b}\} \stackrel{i.i.d.}{\sim} \hat{F}_{Z,n} \quad \theta^{*b} = t(\mathcal{D}^{*b})$$



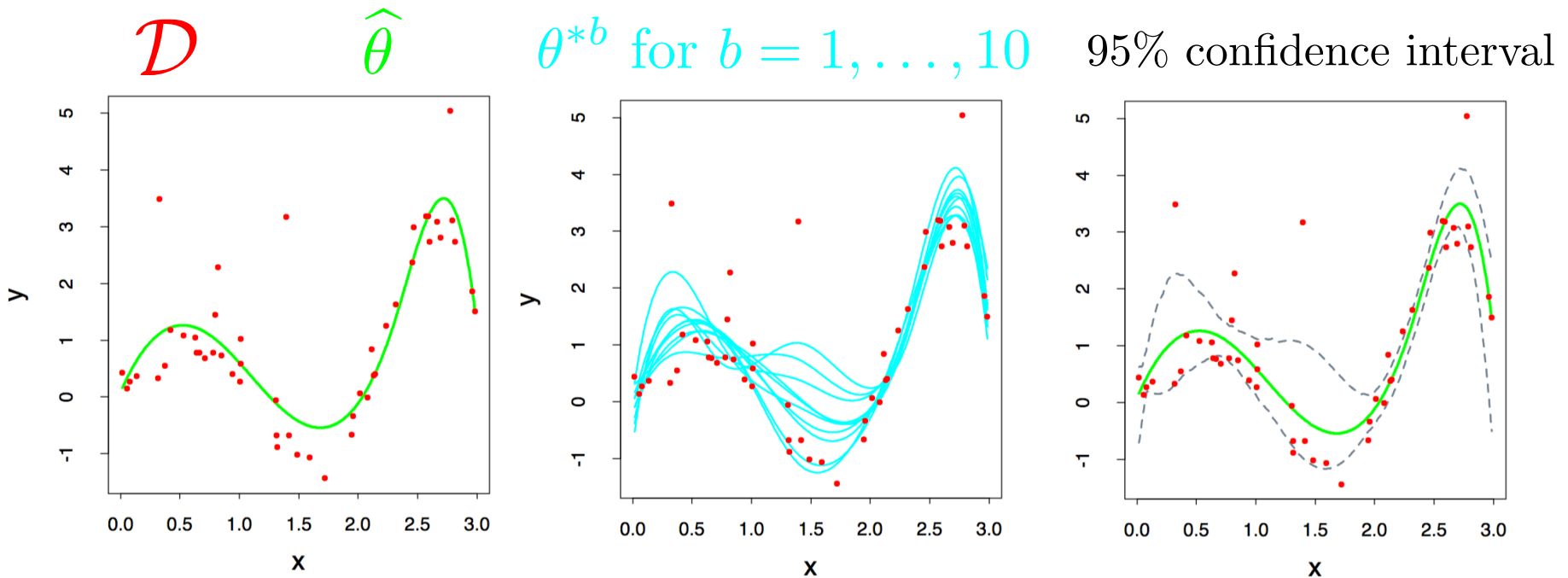
$$|\hat{F}_{Z,n}(x) - F_Z(x)| \stackrel{n \rightarrow \infty}{\rightarrow} 0 \quad \text{a.s.}$$



Applications

Common applications of the bootstrap:

- Estimate parameters that escape simple analysis like the variance or median of an estimate
- Confidence intervals
- Estimates of error for a particular example:



Figures from Hastie et al

Takeaways

Advantages:

- **Bootstrap is very generally applicable. Build a confidence interval around *anything***
- **Very simple to use**
- **Appears to give meaningful results even when the amount of data is very small**
- **Very strong asymptotic theory (as num. examples goes to infinity)**

Takeaways

Advantages:

- Bootstrap is very generally applicable. Build a confidence interval around *anything*
- Very simple to use
- Appears to give meaningful results even when the amount of data is very small
- Very strong asymptotic theory (as num. examples goes to infinity)

Disadvantages

- Very few meaningful finite-sample guarantees
- Potentially computationally intensive
- Reliability relies on test statistic and rate of convergence of empirical CDF to true CDF, which is unknown
- Poor performance on “extreme statistics” (e.g., the max)

Not perfect, but better than nothing.