



MONASH University

ETC3250

Business Analytics

Week 4

The Bootstrap

20 August 2015

Pull yourself up by your bootstraps

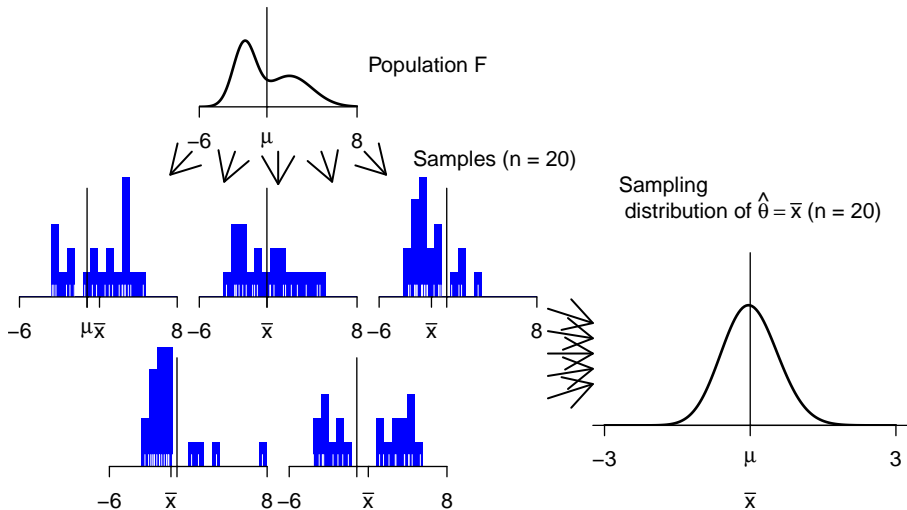


What is the bootstrap?

The bootstrap is a flexible statistical tool to **quantify the uncertainty** associated with a *given estimator or statistical learning method*.

- The bootstrap allows us to use a computer to **mimic the process of obtaining new data sets**, so that we can estimate the **variability of our estimate** without generating additional samples
- We obtain distinct data sets (with the same size as our original dataset) by repeatedly sampling observations **from the original data set with replacement** (nonparametric) or **from an estimated model** (parametric).

Bootstrapping: Ideal world



Bootstrapping: Bootstrap world

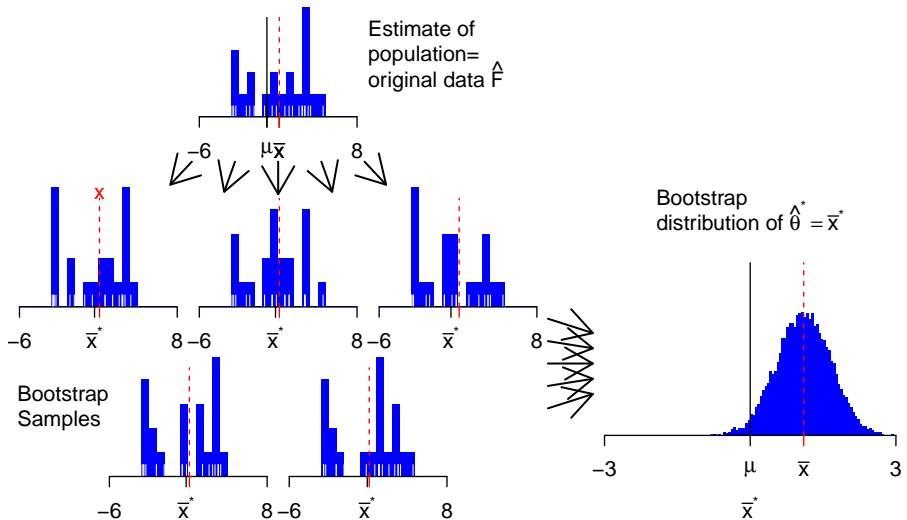
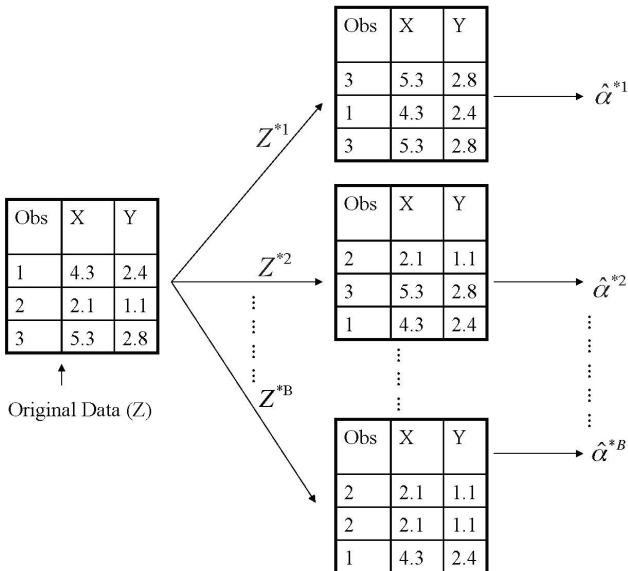


Illustration of the bootstrap



The bootstrap procedure

- Find a good estimate \hat{P} of P
 - Parametric bootstrap
 - Nonparametric bootstrap
- Draw B independent bootstrap samples $X^{*(1)}, \dots, X^{*(B)}$ from \hat{P} :

$$X_1^{*(b)}, \dots, X_n^{*(b)} \sim \hat{P} \quad b = 1, \dots, B.$$

- Evaluate the bootstrap replications:

$$\hat{\theta}^{*(b)} = s(X^{*(b)}) \quad b = 1, \dots, B.$$

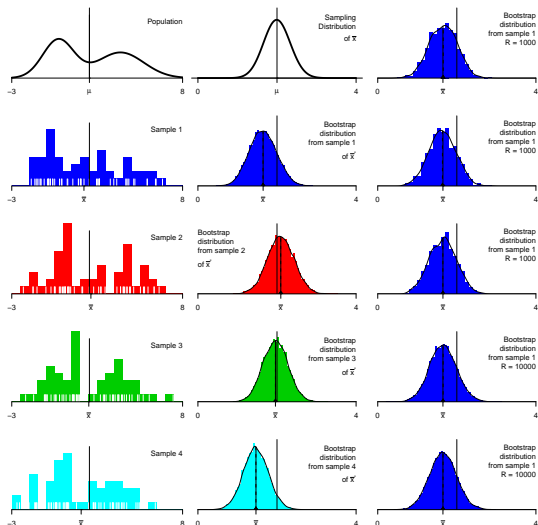
- Estimate the quantity of interest from the distribution of the $\hat{\theta}^{*(b)}$

Examples

What is the standard error of $\hat{\theta}$ (i.e., the standard deviation of the sampling distribution of $\hat{\theta}$)?

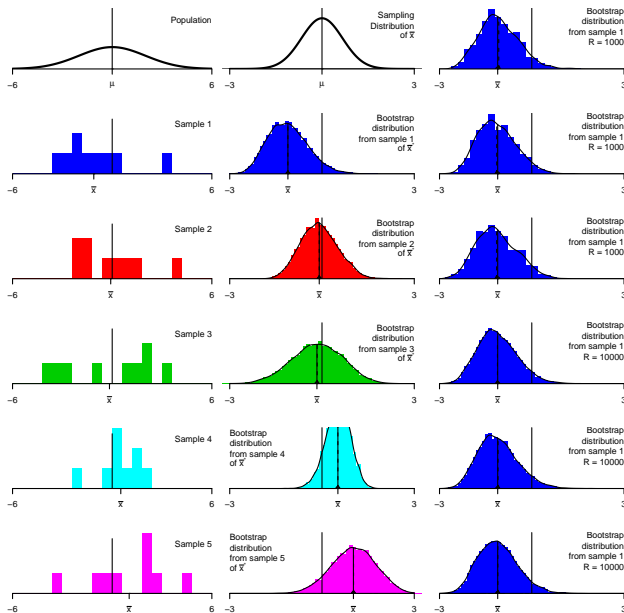
- 1 $\hat{\theta}$ = sample mean
- 2 $\hat{\theta}$ = sample median
- 3 $\hat{\theta}$ = expected shortfall at 5%
- 4 $\hat{\theta}$ = lag 1 autocorrelation.

Sample mean: $n = 50$



■ Two types of random variation

Sample mean: $n = 9$



Prediction error estimation

- Fit the model on a set of bootstrap samples, and then keep track of how well it predicts the original dataset

$$\text{Err}_{\text{boot}} = \frac{1}{B} \frac{1}{N} \sum_{b=1}^B \sum_{i=1}^N L(y_i, \hat{f}^{*b}(x_i))$$

- Each of these bootstrap data sets is created by sampling with replacement, and is the same size as our original dataset. As a result **some observations may appear more than once in a given bootstrap data set and some not at all.**

Prediction error estimation

- Fit the model on a set of bootstrap samples, and then keep track of how well it predicts the original dataset

$$\text{Err}_{\text{boot}} = \frac{1}{B} \frac{1}{N} \sum_{b=1}^B \sum_{i=1}^N L(y_i, \hat{f}^{*b}(x_i))$$

- Each of these bootstrap data sets is created by sampling with replacement, and is the same size as our original dataset. As a result **some observations may appear more than once in a given bootstrap data set and some not at all.**

Prediction error estimation

- Training and validation sets **have observations in common!** Overfit predictions will look very good.

$$P(\text{observation } i \in \text{bootstrap sample } b) = ??$$

Prediction error estimation

- Training and validation sets **have observations in common!** Overfit predictions will look very good.

$$\begin{aligned} P(\text{observation } i \in \text{bootstrap sample } b) &= ?? \\ &= 1 - \left(1 - \frac{1}{n}\right)^n \\ &\approx 1 - \frac{1}{e} \\ &= 0.632 \end{aligned}$$

- Remember that cross-validation uses **non-overlapping** data for the training and validation samples

Prediction error estimation

Better bootstrap version: we only keep track of predictions from bootstrap samples not containing that observation. The leave-one-out bootstrap estimate of prediction error can be defined as

$$\text{Err}_{\text{loo-boot}} = \frac{1}{N} \sum_{i=1}^N \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

where C^{-i} is the set of indices of the bootstrap samples b that do not contain observation i .

Problem of overfitting with Err_{boot} solved but **training-set-size bias as with cross-validation.**

Many applications

- Computing standard errors for complex statistics
- Prediction error estimation
- Bagging (Bootstrap aggregating)
- ...

Variations

There are several types of bootstrap based on different assumptions:

- block bootstrap
- sieve bootstrap
- smooth bootstrap
- residual bootstrap
- wild bootstrap

Many applications

- Computing standard errors for complex statistics
- Prediction error estimation
- Bagging (Bootstrap aggregating)
- ...

Variations

There are several types of bootstrap based on different assumptions:

- block bootstrap
- sieve bootstrap
- smooth bootstrap
- residual bootstrap
- wild bootstrap