Bootstrap

Motivation

Let $x_1, ..., x_n$ be i.i.d samples with distribution F and $x_{(1)}, ..., x_{(n)}$ be the corresponding order statistics . Let $\hat{m}_n = x_{[n/2]}$ be an estimator of the median $m = F^{-1}(1/2)$. How to do inference for m?

It can be showed that

$$\sqrt{n}(\hat{m}_n - m) \to N(0, \frac{1}{4(f(m)^2)})$$

.

- Estimate f(m), then carry out the inference
- Use bootstrap

Bootstrap procedure

Still suppose we have $x_1, ..., x_n$ be i.i.d samples with distribution F. We want to make inference about a statistics $\hat{\theta} = S(x)$ where $x = (x_1, ...x_n)$.

The bootstrap procedure

- 1. Sample $\{x_1^*, ..., x_n^*\}$ with replacement from $\{x_1, ..., x_n\}$.
- 2. Calculate $\hat{\theta}^* = s(x^*)$ where $x = (x_1^*, ...x_n^*)$.
- 3. Repeat 1–2 a total of B times to get $\hat{\theta}_1^*, ..., \hat{\theta}_B^*$, which represents the bootstrap distribution of $\hat{\theta}$.

This sampling approach–sample with replacement from the original dataset is called the empirical bootstrap, invented by Bradley Efron.

Bootstrap procedure

- Bootstrap estimate of variance

$$\hat{Var}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}_b^*)^2$$

where $\bar{\theta}_b^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$.

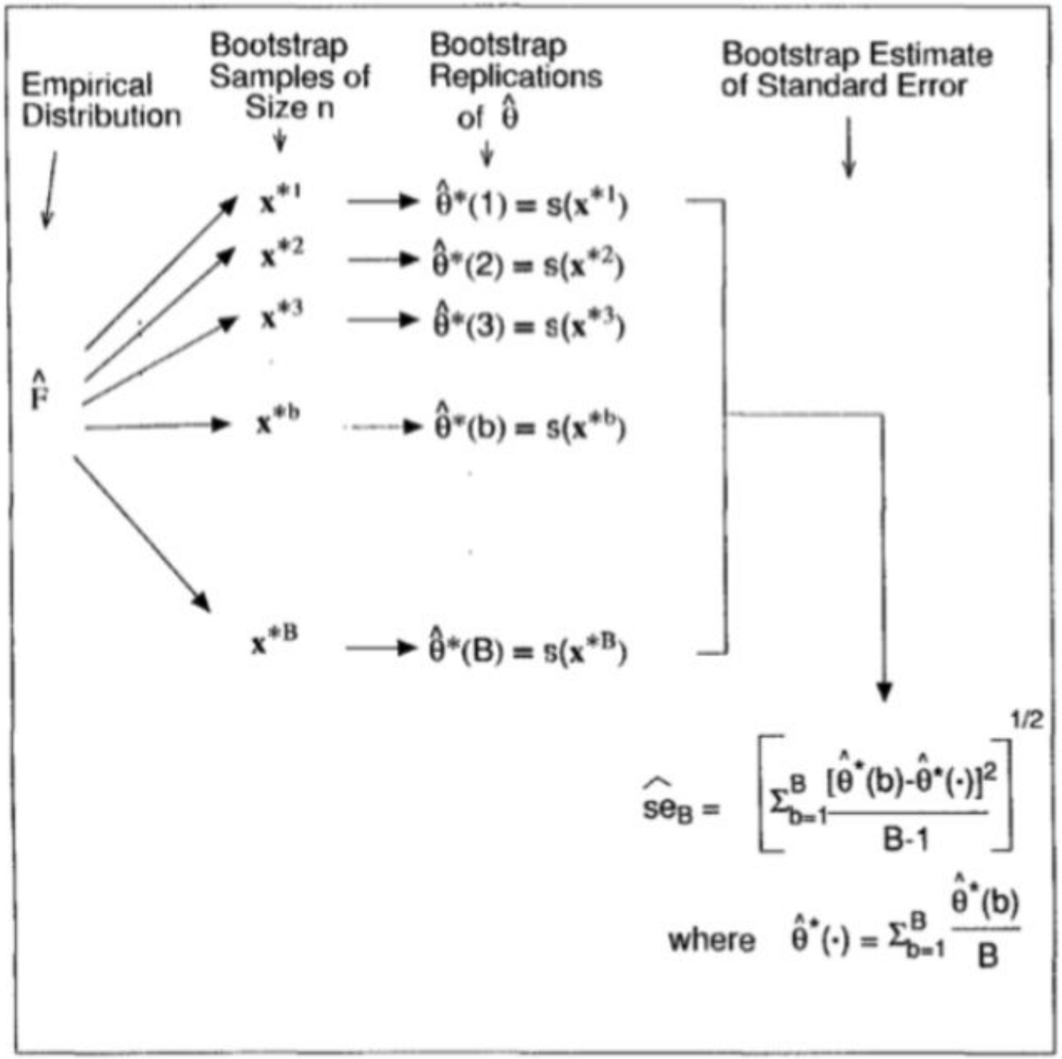
- Bootstrap confidence interval

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\theta})}$$

if $\hat{\theta}$ is asymptotically normal. Or

$$(\hat{\theta}^*_{([\alpha B/2])}, \hat{\theta}^*_{([(1-\alpha/2)B])})$$

Summary so far



Summary so far

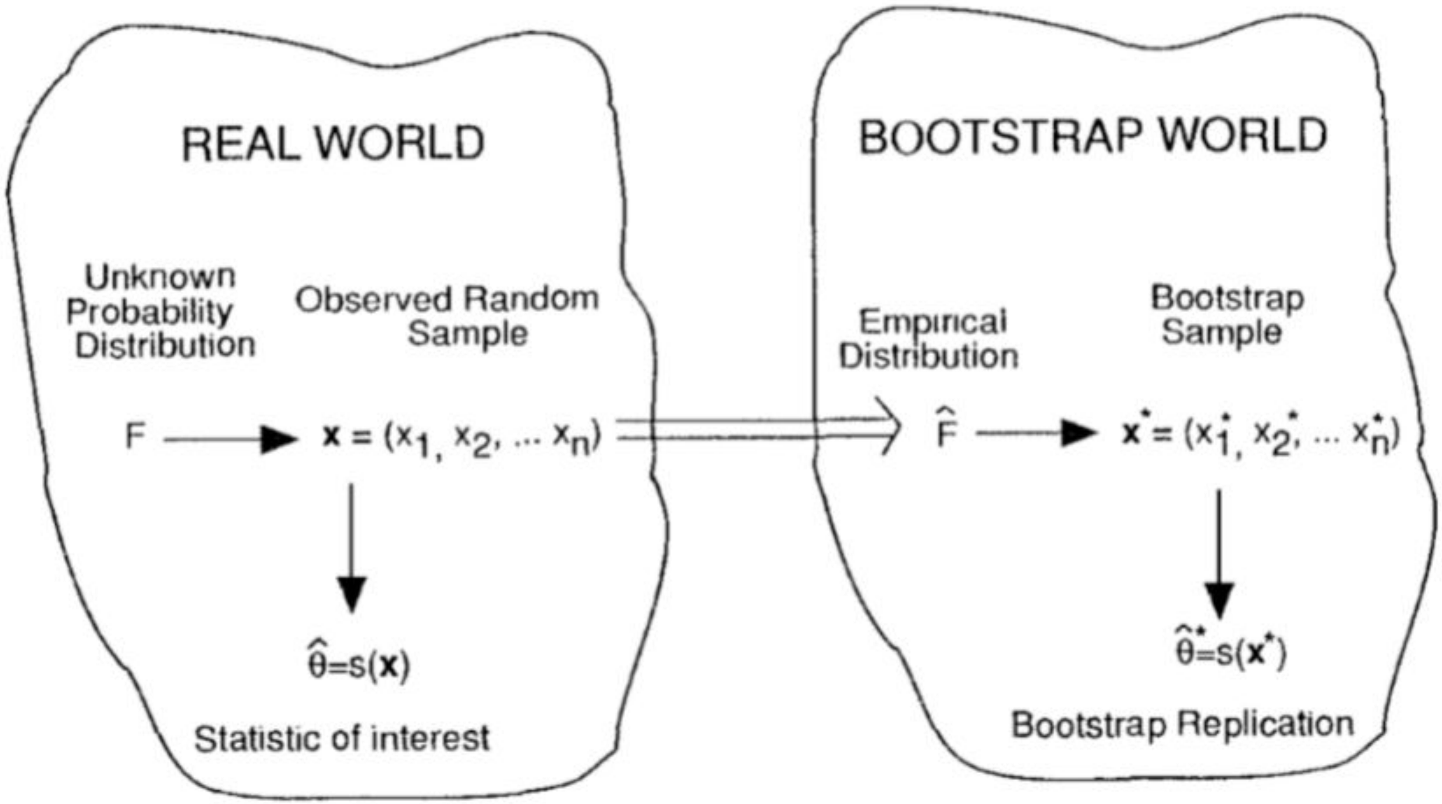


Figure 2: Figure 8.1: An Introduction to the Bootstrap (Efron & Tibshirani, 1993).

Bootstrap in regression

Now consider the following linear model

$$y_i = x_i^T \beta + \epsilon_i$$

Several bootstrap methods are available

- Empirical bootstrap (Paired bootstrap)
- Residual bootstrap

Empirical bootstrap

Direct generalization from the bootstrap for single x into regression setting.

The bootstrap procedure

- 1. Sample $(y_1^*, x_1^*), ..., (y_n^*, x_n^*)$ with replacement from $(y_1, x_1), ..., (y_n, x_n)$.
- 2. Fit the OLS the the bootstrap sample $\{y_i^*, x_i^*\}$ and calculate $\hat{\beta}^*$.
- 3. Repeat 1–2 a total of B times to get bootstrap distribution of $\hat{\beta}^*$.

Empirical bootstrap

- Bootstrap estimate of variance

$$\hat{Var}(\hat{\beta}_j) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}^*_{j,b} - \bar{\beta}^*_j)$$

  where $\bar{\beta}^*_j = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^*_{j,b}$.
- Bootstrap confidence interval

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\beta}_j)}.$$

Or

$$(\hat{\beta}^*_{j,([\alpha B/2])}, \hat{\beta}^*_{j,([(1-\alpha/2)B])})$$

Residual bootstrap

The bootstrap procedure:

- 1. Fit OLS with the original data set and let $\hat{\epsilon}_i = y_i - x_i^T \hat{\beta}$.
- 2. Sample $\epsilon_1^*, ..., \epsilon_n^*$ with replacement from $\hat{\epsilon}_1, ..., \hat{\epsilon}_n$.
- 3. Fit the OLS the the bootstrap sample $\{y_i^*, x_i\}$ where $y_i^* = x_i^T \hat{\beta} + \epsilon_i^*$. Calculate $\hat{\beta}^*$.
- 4. Repeat 1–2 a total of B times to get bootstrap distribution of $\hat{\beta}^*$.

Bootstrap in hypothesis testing

Now consider the following linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Suppose we want to test $H_0 : \beta_1 = \beta_2 = 0$.

Let $F_{obs} = \frac{\sum_i (\hat{y}_i - \bar{y})/2}{\sum_i (y_i - \hat{y}_i)^2/(n-3)}$.

We want to approximate the distribution of F_{obs} from H_0.
Therefore the bootstrap sample should be generated under H_0.

Bootstrap in hypothesis testing

- 1. Let $\hat{\epsilon}_i = y_i - \bar{y}$. Sample $\epsilon_1^*, ..., \epsilon_n^*$ with replacement from $\hat{\epsilon}_1, ..., \hat{\epsilon}_n$.
- 2. Fit y $x_1 + x_2$ with bootstrap sample $\{y_i^*, x_i\}$ where $y_i^* = \bar{y} + \epsilon_i^*$. Calculate $\hat{F}^*$.
- 3. Repeat 1–2 a total of B times to get $\hat{F}_1^*, ..., \hat{F}_B^*$.
- 4. The p-value can be calculated as $\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(F_{obs} > F_b^*)$.