Introduction to Bootstrap

Week 1: bootstrap estimate of standard error

bootstrap은 1979년에 Efron이 처음 제안한 방법으로, 추정치의 표준오차를 구하기 위해서 제안되었다. 여기서 잠깐, 추정치의 표준오차를 구하기 위해 굳이 이런 번거로운 방법을 사용해야할까? 예시를 통해 살펴보자. 우리의 관심사가 데이터의 모평균, μ라고하자. μ를 추정하기 위해 데이터들의 표본평균 $\bar{X}$를 사용할 것이다. n개의 표본을 뽑아서, 실현 값(realization value)을 얻었다고 하자; 이를 $x_1, \cdots, x_n$이라고 한다. 표본 평균이 얼마나 정확한지 알기 위해, 통계학에서는 추정치의 표준오차를 사용하여 신뢰구간을 구하거나 가설검정을 한다. 표본 평균의 표준 오차는 $\frac{\sigma^2}{n}$이다. 물론 여기서 모분산 σ^2가 알려져 있지 않다면 이를 또 추정해서 표준 오차의 추정치를 구해야 한다. 이와 같이 표본 평균이 간단한 형태이고 표준 오차를 유도할 수 있다면 bootstrap을 할 필요가 전혀 없다. 하지만 문제는 추정치의 형태가 이와 같이 간단한 형태가 아닐 때 발생한다. 예를 들어서, 추정량이 $\hat{\theta} = \sum sin(x_i) + 2\sum(x_i - \bar{x})^2$와 같은 형태라면 $\hat{\theta}$의 표준 오차의 closed form을 유도할 수 없다. 바로 이러한 상황에서 boostrap을 이용하여, $se(\hat{\theta})$ 또는 $\hat{se}(\hat{\theta})$을 구하여 $\hat{\theta}$의 uncertainty을 control하는 것이다. 따라서 bootstrap이 소개된 이후, 추정량의 형태가 얼마나 복잡하든, 이것의 표준 오차를 구할 수 있다고 할 수 있다.

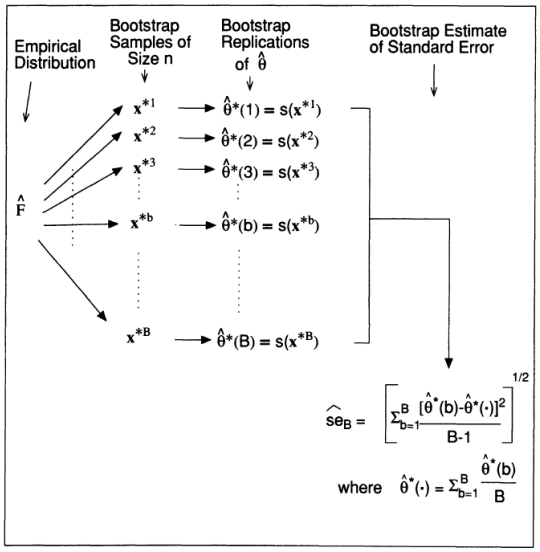
nonparametric bootstrap

분포 가정을 하지 않은 bootstrap을 살펴보자. 우리에게 주어진 데이터가 $x_1, \cdots, x_n$이라고 가정하자. 우리의 관심은 추정량 $\hat{\theta} = s(\mathbf{x})$의 표준 오차를 구하는 것이다. $\hat{F}$를 empirical distribution이라고 하고 각 데이터에 $1/n$의 확률을 부여한다. bootstrap sample을 $\mathbf{x}^* = (x_1^*, \cdots, x_n^*)$이라 하면 $\mathbf{x}^*$은 $\hat{F}$로부터 with replacement으로 뽑힌 n개의 표본이다.

bootstrap sample을 이용하여 $\hat{\theta}^* = s(\mathbf{x}^*)$을 구할 수 있다. 이를 bootstrap replication of $\hat{\theta}$라고 한다. 그런데 우리가 관심있는 것은, bootstrap replication이 아니라, $\hat{\theta}$의 표준 오차이다. 이를 직접적으로 못 구하므로 $\hat{\theta}^*$을 이용하여 간접적으로 구한다. $\hat{\theta}^*$의 표준 오차를 $se_{\hat{F}}(\hat{\theta}^*)$라고 하자. 불행하게도, $se_{\hat{F}}$도 바로 못구하고, 이를 또 추정해야한다. 앞서, 1개의 bootstrap replication을 만들었다. 이제 B개의 bootstrap samples을 이용하여 B개의 bootstrap replications을 만들고, $\hat{\theta}^*$들의 표준 오차를 구한다. 이것이 바로 $se_{\hat{F}}$의 추정치이고, $se_{\hat{F}}$가 바로 $\hat{\theta}^*$의 추정치가 된다. $\hat{\theta}^*$들의 표준 오차를 구할 때, 아래의 표본 표준 오차 공식을 사용한다.

$$\hat{se}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right)^2 \right\}^{1/2}, \; where \; \hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b)$$

아래와 같은 로드맵을 생각하면 더 쉽다.



Parametric bootstrap

parametric boostrap은 bootstrap sample을 뽑는 과정에서 nonparametric bootstrap과 차이를 보인다. 이전에는 empirical distribution을 가정하고, 각 데이터에 $1/n$의 확률을 부여했지만, 이제는 parametric model을 적합하고, 그 모델에서 bootstrap sample을 뽑는다. 예를 들어, 정규분포를 가정하고 평균과 분산의 추정치를 구한 뒤, 이 적합된 분포에서 bootstrap sample을 뽑는다. 그 후에는 동일한 과정을 거친다.

Parametric vs nonparametric bootstrap

직관적으로, nonparametric bootstrap은 분포 가정을 하지 않는 이점이 있다. parametric bootstrap은 분포를 적합하는 수고를 하는 대신에, 더 정확한 결과를 얻을 수 있다고 한다.

아래 그림이 여태까지의 내용을 잘 요약해준다.