전산통계

Chapter 5. 붓스트랩 알고리즘

경험분포함수

- 모수추정에서 모집단 분포에 대한 정보가 없고, 표본 $\{x_1, x_2, \cdots x_n\}$만이 주어져 있는 경우
- 분포함수 $F(x)$에 대한 추정량으로 경험분포함수 (EDF)

$$\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x) \text{ 여기서 } I(x_i \leq x) = \begin{cases} 1, x_i \leq x \\ 0, x_i > x \end{cases}$$

- 임의의 실수 x에 대하여, 경험분포함수는 다음의 성질을 가진다.
  - $n\widehat{F_n}(x) \sim B\big(n, F(x)\big)$
  - $E\left(n\widehat{F_n}(x)\right) = nF(x)$. 즉, $\widehat{F_n}(x)$는 $F(x)$에 대한 불편추정량이다.
  - $Var\left(\widehat{F_n}(x)\right) = \frac{1}{n}F(x)(1 - F(x))$

- 붓스트랩은 분포에 대한 정보없이 자료만 사용한 재표본추출 방법을 통해 통계적 모의실험으로 수행
- 붓스트랩 추정량은 통계량이 아닌 알고리즘의 형태로 표현

- 단계1) 주어진 자료로 EDF를 만든다.

  $x = (x_1, x_2, \cdots x_n)$, 순서통계량: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

- 단계2) 반복단계 (Bootstrap sampling) : $b = 1,2, \cdots, B$
  - 난수생성 $u_i = U(0,1),\ i = 1,2, \cdots n.$
  - 난수와 EDF를 비교하여 새 표본을 생성 (재추출)한다. $u_i, \widehat{F_n}(x) \rightarrow (x_1^{*b}, \cdots, x_n^{*b})$
  - 재추출된 붓스트랩 표본을 사용하여 통계치를 산출한다. $(x_1^{*b}, \cdots, x_n^{*b}) \rightarrow \widehat{\theta^{*b}}$
- 단계3) 구간추정단계 : 산출된 $\widehat{\theta^{*b}},\ b = 1,2, \cdots, B$ 을 이용하여 구간추정을 한다.

구분	기존의 경우	붓스트랩 방법
사용분포	모수가 있는 확률분포 F	경험분포함수 $\hat{F}_n(x)$
자료	관측된 표본	붓스트랩 표본
실험반복	반복이 없음	반복이 있음
추정량	$\hat{\theta} = T(x)$	$\widehat{\theta^*} = T(x^*)$
추정방법	점추정	점추정, 구간추정

- 모평균에 대한 구간추정
- 단계1) $x = (x_1, x_2, \cdots x_n)$으로 경험분포함수를 만든다.
- 단계2) 0과 1사이의 난수 u를 n개 생성한 후 경험분포함수와 비교하여 새로운 표본을 생성한다. 재생성된 붓스트랩 표본으로 추정값 $\overline{x^*}(k)$를 계산한다.
- 단계3) 반복 단계에서 구해진 B개의 붓스트랩 추정값들을 정렬하여 신뢰구간을 구한다.