

经验分布函数: $\{x_1, \dots, x_n\}$ 是从 F 中而来的样本.

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \leq x), \quad -\infty < x < \infty$$

$\mathbb{1}$: 示性函数. $\sum_{i=1}^n \mathbb{1}(x_i \leq x)$ 表示样本中小于等于 x 的个数.

定理: $F_n \xrightarrow{P} F$ 条件: $n \rightarrow \infty$.

证明: $E[\mathbb{1}(x_i \leq x)] = 1 \cdot \Pr(x_i \leq x) + 0 \cdot (1 - \Pr(x_i \leq x))$
 $= \Pr(x_i \leq x)$

$$\begin{aligned} \text{Var}(\mathbb{1}(x_i \leq x)) &= (1 - \Pr(x_i \leq x))^2 \Pr(x_i \leq x) \\ &\quad + (0 - \Pr(x_i \leq x))^2 (1 - \Pr(x_i \leq x)) \\ &= P(1-P) [(1-P) + P] = P(1-P) \end{aligned}$$

$$E\left[\frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \leq x)\right] = \frac{1}{n} n P = P = F(x)$$

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \leq x)\right) = \frac{1}{n^2} n P(1-P) = \frac{1}{n} P(1-P) \rightarrow 0$$

故 $F_n \xrightarrow{m.s.} F$, $F_n \xrightarrow{P} F$

Bootstrap 思想

用样本算统计量, 作为对真实值的估计. 估计误差多大?

例: $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ 作为均值 μ 的估计.

\bar{x}_n 的误差 (标准差) 是 $se(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$. 因为

$$\text{Var}(\bar{x}_n) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$

再用 σ 的估计量, 比如 $\hat{\sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ 代入 $\frac{\sigma}{\sqrt{n}}$, 即可得 \bar{x}_n 的标准差.

但有时事情不这么简单.

任给一统计量 $T_n = g(x_1, \dots, x_n)$, $se(T_n)$ 未必可简单写出.

也即 $\int (T_n - E(T_n))^2 dF(x) \equiv \text{Var}_F(T_n)$ 未知.

用 $\text{Var}_{F_n}(T_n)$ 代替?

想法不错, 但 $\text{Var}_{F_n}(T_n)$ 就能简单写出吗?

我们强行算吧:

○ 从 $\{x_1, \dots, x_n\}$ 中再抽样, 也即再次生成 F_n 的一个抽样分布 $\{x_1^*, \dots, x_n^*\}$.

○ 让 $\{x_1^*, \dots, x_n^*\}$ 生成 B 次, $T_n^* = g(x_1^*, \dots, x_n^*)$ 就有了

$T_{n1}^*, \dots, T_{nB}^*$.

$$\circ \text{Var}_{\text{Boot}} = \frac{1}{B} \sum_{b=1}^B \left(T_{nb}^* - \frac{1}{B} \sum_{r=1}^B T_{nr}^* \right)^2$$

$$\text{Var}_F(T_n) \underset{\text{取决于 } n}{\approx} \text{Var}_{F_n}(T_n) \underset{\text{取决于 } B}{\approx} \text{Var}_{\text{Boot}}(T_n^*)$$