<7장 의문점, 문제>

1. Empirical Distribution Function

Def. of CDF → $F_X(x) = \mathbb{P}(X \le x)$.

7.3 Theorem. *At any fixed value of x,*

$$
\begin{aligned}
\mathbb{E}\left(\widehat{F}_n(x)\right) &= F(x), \\
\mathbb{V}\left(\widehat{F}_n(x)\right) &= \frac{F(x)(1-F(x))}{n}, \\
\text{MSE} &= \frac{F(x)(1-F(x))}{n} \to 0, \\
\widehat{F}_n(x) &\xrightarrow{\text{P}} F(x).
\end{aligned}
$$

7.1 Definition. *The* **empirical distribution function** $\widehat{F}_n$ *is the* CDF *that puts mass $1/n$ at each data point X_i. Formally,*

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \le x)}{n} \tag{7.1}$$

x이하인 점들의 개수 × $\frac{1}{n}$

where

$$I(X_i \le x) = \begin{cases} 1 & \text{if } X_i \le x \\ 0 & \text{if } X_i > x. \end{cases}$$

$n \cdot \hat{F}_n(x) = \sum_{i=1}^{n} I(X_i \le x) \sim \text{Bin}(n, F(x))$

① $E(\hat{F}_n(x)) = F(x)$ ← fixed

$$
\begin{aligned}
&= E\left(\frac{\sum_{i=1}^n I(X_i \le x)}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^n I(X_i \le x)\right) \\
&= \frac{1}{n} \sum_{i=1}^n E[I(X_i \le x)] \\
&= \frac{1}{n} \sum_{i=1}^n P(X_i \le x) \\
&= \frac{1}{n} \sum_{i=1}^n F_{X_i}(x) \\
&= F(x)
\end{aligned}
$$

↲ ?

② $V(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$

$$
\begin{aligned}
&= E[\hat{F}_n(x)^2] - E[\hat{F}_n(x)]^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n P(X_i \le x, X_j \le x) - \frac{1}{n^2} \sum\sum P(X_i \le x) P(X_j \le x) \\
&= \frac{1}{n^2} \sum\sum [P(X_i \le x, X_j \le x) - P(X_i \le x) P(X_j \le x)] \\
&\quad \vdots \\
&= \frac{F(x)(1-F(x))}{n}
\end{aligned}
$$

③

$$\text{MSE} = \frac{F(x)(1-F(x))}{n} \to 0,$$

$$\widehat{F}_n(x) \xrightarrow{\text{P}} F(x).$$

X_n **converges to** X **in probability**, *written* $X_n \xrightarrow{\text{P}} X$, *if, for every* $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \tag{5.1}$$

as $n \to \infty$.

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Chevychev's inequality

$$P(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq \frac{\frac{F(x)(1-F(x))}{n}}{\epsilon^2} = \frac{F(x)(1-F(x))}{n\epsilon^2}$$ 0으로 수렴

$$\therefore \hat{F}_n(x) \xrightarrow{P} F(x)$$

추가정보는 여기에

http://www.win.tue.nl/~rmcastro/AppStat2013/files/lecture1.pdf

9. 100 people are given a standard antibiotic to treat an infection and another 100 are given a new antibiotic. In the first group, 90 people recover; in the second group, 85 people recover. Let p_1 be the probability of recovery under the standard treatment and let p_2 be the probability of recovery under the new treatment. We are interested in estimating $\theta = p_1 - p_2$. Provide an estimate, standard error, an 80 percent confidence interval, and a 95 percent confidence interval for θ.

$\hat{p}_1 = 0.9$
$\hat{p}_2 = 0.85$

$$se(\hat{p}_1) = \sqrt{\frac{V(\hat{p}_1)}{n}} = \sqrt{\frac{0.9 \times 0.1}{100}} = \sqrt{\frac{0.09}{100}}$$

$$se(\hat{p}_2) = \sqrt{\frac{V(\hat{p}_2)}{n}} = \sqrt{\frac{0.85 \times 0.15}{100}} = \sqrt{\frac{0.1275}{100}}$$

① estimate

$$\theta = p_1 - p_2 \Rightarrow \hat{p}_1 - \hat{p}_2 = 0.05$$

② standard error

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) - \underbrace{cov(\hat{p}_1, \hat{p}_2)}_{0}$$

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{Var(\hat{p}_1) + Var(\hat{p}_2)} = 0.04663689 \approx 0.047$$

③ 80 percent interval

$$0.05 \pm 1.28 \times 0.047 \Rightarrow -0.01 \sim 0.11$$

④ 95 percent interval

$$0.05 \pm 2 \times 0.047 \Rightarrow -0.04 \sim 0.14$$