DSL Seminar: 통입+통방 (4)

Kyung-han Kim

Data Science Lab

February, 2023

목차

- 추정, 추정량
  - 점추정
  - 구간추정
- 좋은 추정량이란?
- 표본평균 ($\bar{X}$)과 표본분산 (S^2)
- 표본평균의 평균, 표본평균의 분산
- (범위 외) 표본분산의 평균, 표본분산의 분산
- (다소 범위 외) 중심극한정리

Review

- 연속형 확률변수, 확률분포
- 정규분포
- 이항분포의 정규 근사
- 지수분포
- 균일분포, t분포, F분포, 카이제곱분포

통계적 추정

- 추정 (Estimation): 표본을 바탕으로 모수의 값을 유추하는 일련의 과정.
- 추정량 (Estimator): 모수를 추정하는 식/공식.
- 추정치 (Estimate): 구체적으로 계산된 추정량 값.
- Ex] 대한민국 성인 남성의 키를 유추하기 위해 표본 1,000명을 임의로 추출해 키를 측정하고, 그들의 키의 평균을 구했더니 173cm였다면,
  이 과정 전체를 추정이라고 하고,
  추정량은 표본평균 ($\bar{X}$)을 사용했으며,
  추정치 ($\bar{x}$)는 173cm가 된다고 말할 수 있다.

좋은 추정량의 조건

- 통계학에서는 주로 추정량을 3가지 조건으로 평가합니다:
- 모수가 θ이고 그 추정량이 T_n일 때,
  1. 일치성 (Consistency)
     - $\lim_{n\to\infty} P(|T_n - \theta| > \epsilon) = 0$
  2. 불편성/비편향성 (Unbiasedness)
     - $E[T_n] = \theta$
  3. 효율성 (Efficiency)
     - 모수 θ를 추정하는 T_n이 아닌 추정량 $T_1', T_2', \cdots, T_k'$가 있을 때, $V[T_n] < V[T_i'], \forall i = 1, 2, \cdots, k$
     - 즉, 다른 모든 추정량과 비교했을 때 가장 분산이 작다.
     - 추정량의 분산을 어디까지 줄일 수 있을까? (Rao-Cramer Lower Bound)
     - Unbiased + Greater Variance vs. (Slightly) Biased + Lower Variance?
- 주로 일치성은 충족시키는 경우가 많고, 비편향성을 만족시키는 것을 목표로 합니다. 효율성 또한 중요한 지표이나, 통계학입문에서는 크게 중요하게 다루지는 않습니다. (수리통계학 (2))

점추정과 구간추정

- 점추정 (Point Estimation): 모수의 추정치로 개별 값 하나만을 제시
- 구간추정 (Interval Estimation): 모수의 추정치로 일정 구간을 제시
- Ex] 점추정: 대한민국 성인 남성의 평균 키는 173cm일 것이다.
  구간추정: 대한민국 성인 남성의 평균 키는 170cm $\sim$ 176cm일 것이다.
- 앞으로는 주로 **구간추정**을 다루게 됩니다.

표본평균과 표본분산

- 통계학입문에서 가장 대표적으로 다루는 두 추정량입니다.

$$\text{표본평균: } \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{표본분산: } S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- 표본평균과 표본분산은 비편향 추정량입니다.
  즉, $E[\bar{X}] = \mu$, $E[S^2] = \sigma^2$ 입니다.

cf] 표본평균의 평균?

- 표본이 주어졌을 때, 표본평균은 유일하게 결정됩니다.
- Ex] $x_1 = 100, x_2 = 110, x_3 = 120$이면, $\bar{x} = \frac{x_1+x_2+x_3}{3} = 110$입니다.
- 확률변수의 평균은 의미가 있지만, 상수의 평균은 자기 자신일 뿐이라서 큰 의미가 없습니다.
- 그렇다면 표본평균의 평균을 따진다는 것은 무슨 의미일까요?
- 표본평균은 **확률변수**입니다.
  단, 표본평균은 표본이 바뀔 때 값이 바뀝니다.
  그렇기 때문에 평소에는 표본평균이 확률변수라는 것을 의식하기 어렵습니다.

표본평균의 분산

- 표본평균은 확률변수이기 때문에, 분산도 구할 수 있습니다.

$$V[\bar{X}] = V[\frac{X_1 + X_2 + \cdots + X_n}{n}] = \frac{\sigma^2}{n}$$

- 정규분포를 따르는 확률변수의 합은 정규분포를 따르므로,

$$X \sim N(\mu, \sigma^2) \text{일 때, } \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) = Z$$

- 표준화를 할 때 X는 σ, $\bar{X}$는 σ/n으로 나눈다기보다는,
둘 다 **각자의 표준편차**로 나눈다고 기억하면 좋습니다.

(범위 외) 표본분산의 평균, 표본분산의 분산

- 표본분산도 확률변수이기 때문에, 평균과 분산을 구할 수 있습니다.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2[n-1]$$임을 이용합니다.

- 자유도가 n인 카이제곱분포를 따르는 확률변수의 평균은 n, 분산은 $2n$이므로

$$E[\frac{(n-1)S^2}{\sigma^2}] = n-1,\ V[\frac{(n-1)S^2}{\sigma^2}] = 2(n-1).$$

따라서, $E[S^2] = (n-1)\frac{\sigma^2}{n-1} = \sigma^2,$

$V[S^2] = 2(n-1)\frac{\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}$입니다.

(다소 범위 외) 중심극한정리

- 표본평균은 왜 인기가 많을까?
  1. Unbiased estimator라서?
  2. Minimum-variance estimator라서? (MVUE)
  3. Maximum Likelihood Estimator라서? (MLE)
  4. Intuitive해서?
- 다 맞을 수 있지만, 수학적으로는 중심극한정리가 하나의 이유가 됩니다.
- 중심극한정리: 표본이 충분히 커지면, 표본평균이 모평균의 분포와 상관없이 정규분포에 수렴한다.

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(\mu, \sigma^2)$$

- 즉, X가 어떤 확률분포를 따르는지 모르더라도 $\bar{X}$는 표본이 크면 대략적으로 정규분포를 따르기 때문에 표본평균에 대한 여러 추가적인 분석이 가능해집니다!