풀이

STEP III 2001, 문제 14

확률과 통계 • 분산의 불편추정량

STEP III 2001, 14번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/01-s3-q14

핵심 아이디어

이 문제는 모두 한 가지 사실에 달려 있다. "표본(sample)의 분산(variance)"은 나누는 수로 $n = 2$를 쓰므로 모분산(population variance)을 과소추정하고, 이를 두 배 하면 그 차이가 바로잡힌다. 두 값의 표본평균(sample mean)은 두 값의 중점이므로, 각 값은 표본평균에서 두 값 사이 간격의 절반만큼 떨어져 있다. 따라서 V는 차 $X_1 - X_2$ 하나만의 함수가 되고, 마지막 확률은 정사각형 $[0, a]^2$ 안의 넓이가 된다.

풀이

X의 분산

X의 확률밀도함수(density)는 $[0, a]$에서 $f(x) = \frac{1}{a}$이다. 대칭성(symmetry)에 의해 $\mathrm{E}(X) = \frac{1}{2}a$이고,

$$\mathrm{E}(X^2) = \int_0^a \frac{x^2}{a}\,\mathrm{d}x = \frac{a^2}{3}.$$

따라서

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12},$$

이므로 요구된 결과를 얻는다.

표본분산

표본 X_1, X_2의 표본평균은 $\bar{X} = \frac{1}{2}(X_1 + X_2)$이고, 표본의 분산은 표본평균으로부터의 편차제곱의 평균이다.

$$V = \tfrac{1}{2}[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2].$$

이제 $X_1 - \bar{X} = \frac{1}{2}(X_1 - X_2)$이고 $X_2 - \bar{X} = -\frac{1}{2}(X_1 - X_2)$이므로, 두 편차 제곱은 모두 $\frac{1}{4}(X_1 - X_2)^2$과 같다. 그러므로

$$V = \tfrac{1}{2} \cdot 2 \cdot \tfrac{1}{4}(X_1 - X_2)^2 = \tfrac{1}{4}(X_1 - X_2)^2,$$

이고, 이것이 주어진 형태이다.

생각해 보기 이중적분 없이 $\mathrm{E}((X_1 - X_2)^2)$를 어떻게 구할까?

X_1과 X_2는 같은 분포를 따르므로 $\mathrm{E}(X_1 - X_2) = 0$이고, 따라서 $\mathrm{E}((X_1 - X_2)^2) = \mathrm{Var}(X_1 - X_2)$이다. X_1과 X_2는 독립(independent)이므로 분산끼리 더해진다($\mathrm{Var}(-X_2) = \mathrm{Var}(X_2)$이므로 빼기 부호는 상관없다).

$$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \frac{a^2}{12} + \frac{a^2}{12} = \frac{a^2}{6}.$$

그러므로

$$\mathrm{E}(2V) = \tfrac{1}{2}\,\mathrm{E}((X_1 - X_2)^2) = \tfrac{1}{2} \cdot \frac{a^2}{6} = \frac{a^2}{12} = \mathrm{Var}(X).$$

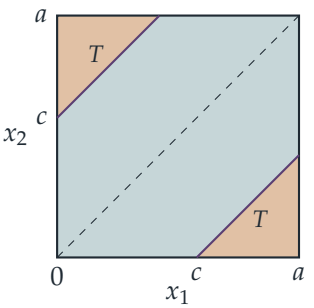
a의 값이 무엇이든 $2V$의 기댓값(expected value)은 $2V$가 추정하는 양과 같으므로, $2V$는 X의 분산의 불편추정량(unbiased estimator)이다. (이와 달리 $\mathrm{E}(V) = \frac{a^2}{24}$로, 참값인 분산의 절반밖에 되지 않는다.)

확률

$V \geq 0$이고 $\frac{1}{12}a^2 > 0$이므로 제곱근을 취할 수 있다.

$$V < \frac{a^2}{12} \iff (X_1 - X_2)^2 < \frac{a^2}{3} \iff |X_1 - X_2| < \frac{a}{\sqrt{3}}.$$

X_1과 X_2는 독립이고 각각 균등분포(uniform distribution)를 따르므로, 순서쌍 (X_1, X_2)는 정사각형 $[0, a]^2$ 위에서 균등분포를 따른다. 따라서 구하는 확률은 $c = a/\sqrt{3} < a$라 할 때 정사각형의 넓이 가운데 띠 영역 $|x_1 - x_2| < c$에 놓인 부분의 비율이다.



여사건(complement)은 색칠된 두 모퉁이 삼각형 T로 이루어지며, 각각은 빗변이 아닌 두 변의 길이가 $a-c$인 직각이등변삼각형(right-angled isosceles triangle)이다. 두 삼각형의 넓이의 합은 $2\cdot\frac{1}{2}(a-c)^2=(a-c)^2$이므로

$$\mathrm{P}\left(V<\frac{a^2}{12}\right)=1-\frac{(a-c)^2}{a^2}=1-\left(1-\frac{1}{\sqrt{3}}\right)^2=1-\left(\frac{4}{3}-\frac{2}{\sqrt{3}}\right)=\frac{2}{\sqrt{3}}-\frac{1}{3}.$$

즉, 정확한 확률은

$$\mathrm{P}\left(V<\frac{a^2}{12}\right)=\frac{2\sqrt{3}-1}{3}.$$

$\sqrt{3}\approx 1.732$를 쓰면 이 값은 $\frac{2.464}{3}\approx 0.821$이므로, 유효숫자(significant figure) 한 자리로 나타낸 확률은 0.8이다.

검산

- 답은 a에 의존하지 않으며, 마땅히 그래야 한다. X를 어떤 배율로 바꾸면 V와 $\frac{1}{12}a^2$가 모두 그 배율의 제곱만큼 바뀌기 때문이다.
- $a=3$으로 4×10^5개의 순서쌍을 시뮬레이션한 결과, $2V$의 평균은 0.751(목표값 0.75)이었고 $V<\frac{1}{12}a^2$인 비율은 0.821로 $\frac{1}{3}(2\sqrt{3}-1)=0.8214$와 잘 맞았다.
- 타당성 점검: 확률은 $(0,1)$ 안에 있고, $\frac{1}{12}a^2$를 $\frac{1}{4}a^2$(V가 가질 수 있는 가장 큰 값)로 바꾸면 $c=a$가 되어 공식 $1-(1-c/a)^2$이 1을 준다.

흔한 실수

- "표본의 분산"에 나누는 수로 $n-1$을 쓰는 것: 그러면 주어진 $\frac{1}{4}(X_1-X_2)^2$이 아니라 $\frac{1}{2}(X_1-X_2)^2$이 나온다. 이 문제는 나누는 수로 n을 쓰기를 요구한다.
- 절댓값(absolute value)을 빠뜨려서 두 삼각형 가운데 하나만 세는 것(그러면 확률이 $\frac{1}{2}$보다 작게 나온다).
- $\mathrm{Var}(X_1-X_2)=\mathrm{Var}(X_1)-\mathrm{Var}(X_2)=0$이라고 쓰는 것.

STEP 너머

여기서 얻은 결과는 일반적인 정리의 $n=2$인 경우이다. 분산이 σ^2인 분포에서 뽑은 크기 n인 임의의 표본에 대하여 $\sum(X_i-\bar{X})^2$의 평균은 $(n-1)\sigma^2$이다. 나누는 수로 $n-1$을 쓰면(베셀 보정(Bessel's correction)) 불편추정량이 되는 것은 이 때문이다. 대학 1학년 통계학에서는 더 나아가 불편추정량 가운데 어느 것의 분산이 가장 작은지를 묻는다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문), 그리고 Blitzstein과

Hwang의 *Introduction to Probability*(확률론 입문)는 모두 여기서 쓴 방식으로 합의 기댓값과 분산을 다룬다.

Beyond STEP 풀이: https://beyondstep.keithmcnulty.org/ko/. 풀이 글은 CC0 1.0이다. 문제 자체의 저작권은 © Cambridge University Press & Assessment에 있으며 여기에 싣지 않는다. 영어 원문을 기계 번역하였으며 아직 검토되지 않았다.