풀이

STEP II 1996, 문제 14

확률과 통계 • 적률과 중심극한정리

> **STEP II 1996, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/96-s2-q14

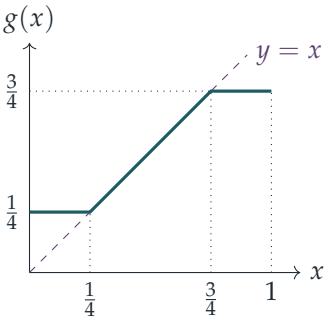
핵심 아이디어

Y는 X의 양 끝을 "잘라낸" 것이다. 즉 $\frac{1}{4}$ 아래의 확률은 모두 한 값 $\frac{1}{4}$에 쌓이고, $\frac{3}{4}$ 위의 확률은 모두 $\frac{3}{4}$에 쌓인다. 따라서 Y는 두 개의 점질량(point mass)과 연속인 가운데 부분으로 이루어진 혼합 확률변수(mixed random variable)이고, 모든 기댓값(expectation) $\mathrm{E}(g(Y))$도 같은 방식으로 세 부분으로 나뉜다. 마지막 부분에서는 서로 독립(independent)인 매우 많은 복사본의 합을 다루므로 중심극한정리(central limit theorem)가 필요하다. 이를 위해서는 Y의 평균(mean)과 분산(variance)만 있으면 되는데, 앞 부분에서 이미 이 작업을 해 두었다.

풀이

Y의 적률(moments)

$X \sim \mathrm{U}[0,1]$이므로 $\mathrm{P}(X \le \frac{1}{4}) = \mathrm{P}(X \ge \frac{3}{4}) = \frac{1}{4}$이고, $[\frac{1}{4}, \frac{3}{4}]$에서 X의 확률밀도(density)는 1이다. 아래 그림의 자르기 함수 g를 써서 $Y = g(X)$로 나타내면 $\mathrm{E}(Y^n) = \int_0^1 g(x)^n \,\mathrm{d}x$를 얻는다.



적분을 $\frac{1}{4}$과 $\frac{3}{4}$에서 나누면

$$\begin{aligned}\mathrm{E}(Y^n) &= \left(\tfrac{1}{4}\right)^n \cdot \tfrac{1}{4} + \int_{1/4}^{3/4} x^n \,\mathrm{d}x + \left(\tfrac{3}{4}\right)^n \cdot \tfrac{1}{4} \\ &= \frac{1}{4^{n+1}} + \frac{3^{n+1}-1}{(n+1)4^{n+1}} + \frac{3^n}{4^{n+1}} \\ &= \frac{(n+1)(1+3^n)+3^{n+1}-1}{(n+1)4^{n+1}}.\end{aligned}$$

분자는 $n+1+(n+1)3^n+3\cdot 3^n-1 = n+(n+4)3^n$이므로

$$\boxed{\mathrm{E}(Y^n) = \frac{n+(n+4)\,3^n}{(n+1)\,4^{n+1}}} \qquad (n \geq 1).$$

X와 비교하기

균등분포(uniform distribution)를 따르는 확률변수에 대해서는 $\mathrm{E}(X^n) = \int_0^1 x^n \,\mathrm{d}x = \frac{1}{n+1}$이다.

$n=1$이면 공식에서 $\mathrm{E}(Y) = \frac{1+5\cdot 3}{2\cdot 16} = \frac{1}{2} = \mathrm{E}(X)$를 얻으며, 이것이 보이려던 것이다. (이는 놀라운 일이 아니다. 자르기는 양 끝의 질량을 $\frac{1}{2}$에 대하여 대칭적으로, 같은 만큼씩 안쪽으로 옮기기 때문이다.)

$n=2$이면 $\mathrm{E}(Y^2) = \frac{2+6\cdot 9}{3\cdot 64} = \frac{56}{192} = \frac{7}{24}$이고, 한편 $\mathrm{E}(X^2) = \frac{1}{3} = \frac{8}{24}$이다. 따라서

$$\mathrm{E}(X^2) - \mathrm{E}(Y^2) = \frac{8}{24} - \frac{7}{24} = \frac{1}{24},$$

이며, 이것이 보이려던 것이다.

일반적인 $n \geq 2$에 대해서는 두 적률을 공통분모 $(n+1)4^{n+1}$로 나타낸다.

$$\mathrm{E}(X^n) = \frac{4^{n+1}}{(n+1)4^{n+1}}.$$

따라서 $4^{n+1} > n + (n+4)3^n$임을 보이면 된다. $4^{n+1} = 4(3+1)^n$과 이항정리(binomial theorem)를 쓰고, 위의 두 항을 따로 떼어 두면

$$4^{n+1} = 4 \cdot 3^n + 4n \cdot 3^{n-1} + 4\sum_{k=0}^{n-2} \binom{n}{k} 3^k.$$

여기서 $n + (n+4)3^n = 4 \cdot 3^n + 3n \cdot 3^{n-1} + n$을 빼면

$$4^{n+1} - (n + (n+4)3^n) = n\,3^{n-1} - n + 4\sum_{k=0}^{n-2} \binom{n}{k} 3^k.$$

? **생각해 보기** 이 차는 왜 양수인가?

$n \geq 2$이면 $3^{n-1} \geq 3$이므로 $n\,3^{n-1} - n \geq 2n > 0$이다. 합 또한 양수인데, 1과 같은 $k = 0$인 항을 포함하고 모든 항이 음이 아니기 때문이다. 따라서 차는 엄밀하게 양수이고

$$\mathrm{E}(X^n) > \mathrm{E}(Y^n) \qquad \text{모든} n \geq 2.$$

이다. ($n = 1$이면 차는 $3^0 - 1 + 0 = 0$이고, 이는 $\mathrm{E}(X) = \mathrm{E}(Y)$와 일치한다.)

240 000개 복사본의 합

앞 부분으로부터

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \frac{7}{24} - \frac{1}{4} = \frac{1}{24}.$$

$N = 240\,000$일 때 $S = Y_1 + \cdots + Y_N$이라 하자. Y_i들은 서로 독립이므로 기댓값과 분산이 각각 더해진다.

$$\mathrm{E}(S) = \frac{N}{2} = 120\,000, \qquad \mathrm{Var}(S) = \frac{N}{24} = 10\,000,$$

따라서 S의 표준편차(standard deviation)는 100이다. Y_i들은 서로 독립이고 분산이 유한한 같은 분포를 따르며 N이 매우 크므로, 중심극한정리에 의해 S는 근사적으로 $\mathrm{N}(120\,000,\, 100^2)$를 따른다. 따라서

$$\mathrm{P}(S < K) \approx \Phi\Big(\frac{K - 120\,000}{100}\Big) = \frac{3}{4} \quad \Longrightarrow \quad \frac{K - 120\,000}{100} = \Phi^{-1}(0.75) \approx 0.6745,$$

이므로

$$K \approx 120\,000 + 67.45 \approx 120\,067.$$

연속성 수정(continuity correction)은 필요하지 않다. Y는 정숫값을 갖지 않으며, 어차피 표준편차가 100이므로 그런 수정은 이 정도의 정확도에서는 아무런 차이를 만들지 않는다.

검산

- $\frac{1}{4} \leq Y \leq \frac{3}{4}$이므로 $\mathrm{E}(Y^n)$의 공식은 당연히 $(\frac{1}{4})^n$과 $(\frac{3}{4})^n$ 사이에 있어야 하며, 실제로 그렇다. $n = 3$이면 $\frac{3+7\cdot 27}{4\cdot 256} = \frac{192}{1024} = \frac{3}{16} = 0.1875$를 얻고, $g(x)^3$을 중점법(midpoint rule)으로 적분한 값과 소수점 아래 열째 자리까지 일치한다.

- $\mathrm{Var}(Y) = \frac{1}{24} < \frac{1}{12} = \mathrm{Var}(X)$이다. 자르기는 값들을 평균 쪽으로 끌어당기므로 퍼짐이 줄어드는 것이 맞다.
- Y의 복사본 2400개의 합을 20 000번 시뮬레이션하였더니 표준화된 제3사분위수(upper quartile)가 약 0.678로 0.6745에 가까웠으며, 이는 정규근사(normal approximation)를 뒷받침한다.

흔한 실수

- Y를 $[\frac{1}{4}, \frac{3}{4}]$에서의 균등분포로 보고, 양 끝에 있는 크기 $\frac{1}{4}$의 점질량을 잊는 것. 그 구간에서의 균등분포를 따르는 확률변수의 확률밀도는 1이 아니라 2이다.
- $N\,\mathrm{Var}(Y)$ 대신 $\mathrm{Var}(S) = N^2\,\mathrm{Var}(Y)$를 쓰는 것. 독립이면 분산이 더해지므로 표준편차는 $\sqrt{N}$처럼 커진다.
- 마지막 단계에서 부호를 틀리는 것. $\mathrm{P}(S < K) = \frac{3}{4} > \frac{1}{2}$이므로 K는 평균보다 커야 한다.

STEP 너머

Y는 X를 자른(또는 "윈저화(Winsorised)"한) 것이며, 이산도 연속도 아닌 분포를 가진다. 측도론적 확률론(measure-theoretic probability)은 이런 분포를 분포함수(distribution function)를 통해 일관되게 다루며, 모든 기댓값은 하나의 적분 $\int g\,\mathrm{d}F$가 된다. 평균이 같을 때의 부등식 $\mathrm{E}(X^n) > \mathrm{E}(Y^n)$은 볼록 순서(convex ordering)의 한 예이다. X는 Y의 "평균 보존 확산(mean-preserving spread)"이므로, 모든 볼록(convex) 함수 ϕ에 대하여 $\mathrm{E}(\phi(X)) \geq \mathrm{E}(\phi(Y))$이다. 마지막에 쓴 중심극한정리는 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)에 증명되어 있다.

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.