풀이

STEP I 2005, 문제 14

확률과 통계 • 혼합형 확률변수

> **STEP I 2005, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/05-s1-q14

핵심 아이디어

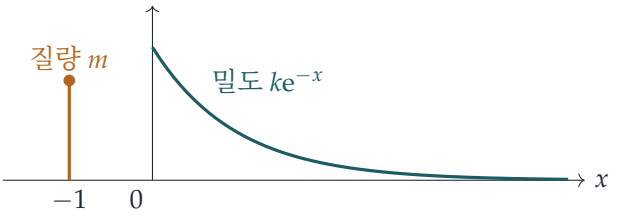
X는 **혼합** 확률변수(mixed random variable)이다. 즉, 한 점 -1에 확률이 덩어리로 몰려 있고, 이와 함께 $[0,\infty)$ 위에 연속적으로 퍼져 있는 부분이 있다. 이런 확률변수를 다루려면 모든 계산을 두 부분으로 나누면 된다. $X=-1$에서 오는 이산 항에, $[0,\infty)$ 위의 확률밀도함수에 대한 적분을 더하는 것이다. 연속 부분이 단지 상수배된 지수분포(exponential distribution)라는 사실을 알아차리면, 대부분의 적분은 이미 알고 있는 것들이다.

풀이

주어진 누적확률을 미분하면, X의 연속 부분의 확률밀도함수(density)는

$$f(x)=\frac{\mathrm{d}}{\mathrm{d}x}k(1-\mathrm{e}^{-x})=k\mathrm{e}^{-x},\qquad x\geq 0.$$

이다. 따라서 X는 다음과 같은 모양이다. -1에 점질량(point mass) m이 있고, $[0,\infty)$ 위에 곡선 $k\mathrm{e}^{-x}$가 있다.



소문항 (i)

전체 확률은 1이다. $x \to \infty$로 보내면 $\mathrm{P}(X \geq 0) = k(1-0) = k$이므로

$$m + k = 1, \qquad \text{즉,} \qquad k = 1 - m.$$

이다. ($m \geq 0$은 확률이고 $m < \frac{1}{2}$이므로 $k > \frac{1}{2} > 0$이다. 따라서 밀도는 실제로 양수이다.)

소문항 (ii)

두 부분으로 나누면

$$\begin{aligned}\mathrm{E}(X) &= (-1)\cdot m + \int_0^\infty x\,(1-m)\mathrm{e}^{-x}\,\mathrm{d}x \\ &= -m + (1-m)\cdot 1 = 1 - 2m,\end{aligned}$$

이다. 여기서 $\int_0^\infty x\mathrm{e}^{-x}\,\mathrm{d}x = 1$을 이용하였다(부분적분(integration by parts)을 하거나, 모수가 1인 지수분포의 평균임을 알아보면 된다). 이것이 구하려던 결과이다.

소문항 (iii)

분산(*variance*). 같은 방법으로, $\int_0^\infty x^2\mathrm{e}^{-x}\,\mathrm{d}x = 2$(부분적분을 두 번 한다)를 이용하면

$$\mathrm{E}(X^2) = (-1)^2\cdot m + (1-m)\cdot 2 = 2 - m.$$

이다. 따라서

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 2 - m - (1-2m)^2 \\ &= 1 + 3m - 4m^2 = (1-m)(1+4m).\end{aligned}$$

생각해 보기 중앙값이 -1일 수 있을까? 아니면 -1과 0 사이 어딘가에 있을 수 있을까?

중앙값(*median*). 그럴 수 없다. X는 $(-1, 0)$의 값을 결코 갖지 않으며, $\mathrm{P}(X \leq -1) = m < \frac{1}{2}$이므로 점질량만으로는 확률의 절반에 이르지 못한다. 바로 이 부분에서 조건 $m < \frac{1}{2}$가 쓰인다. 따라서 중앙값 M은 $M \geq 0$이고 $\mathrm{P}(X \leq M) = \frac{1}{2}$를 만족한다.

$$\begin{aligned}m + (1-m)(1-\mathrm{e}^{-M}) &= \tfrac{1}{2} \\ 1 - (1-m)\mathrm{e}^{-M} &= \tfrac{1}{2} \\ \mathrm{e}^{-M} &= \frac{1}{2(1-m)}.\end{aligned}$$

그러므로

$$M = \ln\left(2(1-m)\right) = \ln(2-2m).$$

이다. 앞뒤가 맞는지 확인해 보면, $0 \le m < \frac{1}{2}$에서 $1 < 2 - 2m \le 2$이므로 $M > 0$이고, 마땅히 그래야 한다. 분포가 $[0, \infty)$에서 연속이므로 이 M은 $\mathrm{P}(X \ge M) = \frac{1}{2}$도 만족하고, 따라서 중앙값은 하나로 정해진다.

소문항 (iv)

$X = -1$일 때 $|X|^{1/2} = 1$이고, $X \ge 0$일 때 $|X|^{1/2} = \sqrt{X}$이다. 따라서

$$\mathrm{E}(|X|^{1/2}) = 1 \cdot m + (1 - m) \int_0^\infty \sqrt{x}\, \mathrm{e}^{-x}\, \mathrm{d}x.$$

이다.

? **생각해 보기** $\int_0^\infty \sqrt{x}\, \mathrm{e}^{-x}\, \mathrm{d}x$를 어떻게 주어진 적분으로 바꿀 수 있을까?

주어진 적분에는 e^{-y^2}이 있으므로 $x = y^2$ $(y \ge 0)$으로 치환하는 것이 자연스럽다. 그러면 $\sqrt{x} = y$, $\mathrm{d}x = 2y\, \mathrm{d}y$이고, 적분 구간의 끝 0과 ∞는 그대로이다.

$$\int_0^\infty \sqrt{x}\, \mathrm{e}^{-x}\, \mathrm{d}x = \int_0^\infty y\, \mathrm{e}^{-y^2}\, 2y\, \mathrm{d}y = 2 \int_0^\infty y^2 \mathrm{e}^{-y^2}\, \mathrm{d}y = 2 \cdot \tfrac{1}{4}\sqrt{\pi} = \tfrac{1}{2}\sqrt{\pi}.$$

따라서

$$\mathrm{E}(|X|^{1/2}) = m + \tfrac{1}{2}\sqrt{\pi}\,(1 - m).$$

이다.

검산

- $m = 0$이면 X는 정확히 모수가 1인 지수분포를 따른다. 위의 공식들은 평균 1, 분산 1, 중앙값 $\ln 2$, $\mathrm{E}(\sqrt{X}) = \frac{1}{2}\sqrt{\pi} = \Gamma(\frac{3}{2})$를 주며, 모두 잘 알려진 값이다.
- $m \to \frac{1}{2}$이면 중앙값은 $\ln 1 = 0$으로 수렴하는데, 이는 이치에 맞다. 확률의 절반이 -1에 있으므로 중앙값은 연속 부분의 맨 아래 끝까지 미끄러져 내려간다.
- $m = 0.3$으로 두고 4×10^5개의 표본으로 모의실험(simulation)을 한 결과, 평균 0.402(공식 0.4), 분산 1.551(공식 1.54), 중앙값 0.336(공식 $\ln 1.4 \approx 0.336$), $\mathrm{E}(|X|^{1/2}) \approx 0.922$(공식 0.920)를 얻었다.

흔한 실수

- $\mathrm{E}(X^2)$나 $\mathrm{E}(|X|^{1/2})$에서 점질량을 빠뜨리는 것. $(-1)^2 = |-1|^{1/2} = 1$이므로 점질량은 $-m$이 아니라 $+m$만큼 기여한다.

- 중앙값을 구할 때 $(1-m)(1-\mathrm{e}^{-M}) = \frac{1}{2}$를 푸는 것. 이렇게 하면 -1에 이미 쌓여 있는 확률 m을 무시하게 된다.
- 치환할 때 $\mathrm{d}x$를 $2y\,\mathrm{d}y$로 바꾸지 않는 것. 그러면 $\frac{1}{2}\sqrt{\pi}$ 대신 $\frac{1}{4}\sqrt{\pi}$가 나온다.

STEP 너머

점질량과 밀도가 섞인 분포는 실제로 흔하다(예를 들어 보험 청구액은 어떤 확률로 0이고, 그렇지 않으면 연속적인 값을 갖는다). 대학에서는 기댓값을 분포함수(distribution function)에 대한 적분 $\mathrm{E}(g(X)) = \int g(x)\,\mathrm{d}F(x)$로 써서 이런 분포를 한꺼번에 다루는데, 이 방법은 이산, 연속, 혼합인 경우를 모두 아우른다. 소문항 (iv)의 적분은 감마함수(Gamma function)의 값 $\Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{\pi}$이다. 다음 단계로는 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)이 좋다.

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.