풀이

STEP II 2009, 문제 12

확률과 통계 • 반정규분포

> **STEP II 2009, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/09-s2-q12

핵심 아이디어

지수 $-2x^2$는 모습을 감춘 정규분포(normal distribution)의 밀도이다. $-2x^2 = -\dfrac{x^2}{2\sigma^2}$로 쓰면 $\sigma = \frac{1}{2}$를 얻으므로, f는 $Y \sim \mathrm{N}(0, \frac{1}{4})$의 확률밀도함수(probability density function)를 0에서 잘라 낸 것의 정확히 두 배이다. 다시 말해 X는 $|Y|$와 같은 분포를 가진다. 이것을 알아차리고 나면 정규화 상수(normalising constant), $\mathrm{E}(X^2)$, 그리고 중앙값(median)이 모두 정규분포에 관한 사실들로부터 곧바로 나온다. $\mathrm{E}(X)$만이 (아주 쉬운) 적분을 필요로 한다.

풀이

이 풀이 전체에서 $Y \sim \mathrm{N}(0, \frac{1}{4})$, $Z \sim \mathrm{N}(0,1)$이라 하자. 그러면 Y는 $\frac{1}{2}Z$와 같은 분포를 가지며, 그 밀도는 다음과 같다.

$$g(y) = \frac{1}{\frac{1}{2}\sqrt{2\pi}}\,\mathrm{e}^{-y^2/(2\cdot\frac{1}{4})} = \frac{2}{\sqrt{2\pi}}\,\mathrm{e}^{-2y^2}.$$

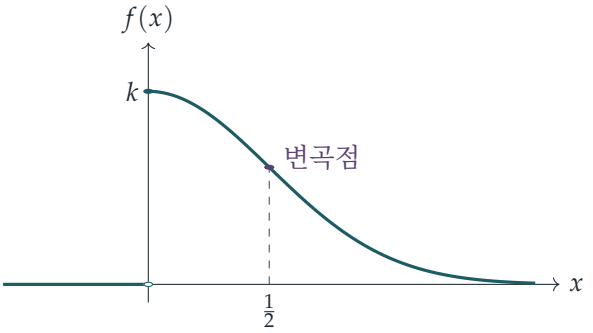
소문항 (i)

$x \geq 0$에서 $f'(x) = -4kx\,\mathrm{e}^{-2x^2} \leq 0$이므로, f는 $x = 0$에서의 최댓값 k로부터 감소하며, 그 점에서 기울기는 영이다. 또한

$$f''(x) = k(16x^2 - 4)\,\mathrm{e}^{-2x^2},$$

이고, 이는 $x = \frac{1}{2}$에서 부호가 바뀌므로 $x = \frac{1}{2}$에 변곡점(point of inflection)이 있다. $x \to \infty$일 때 $f(x) \to 0$으로 매우 빠르게 수렴하며, $x < 0$에서는 $f(x) = 0$이다. 그래프는

종 모양 곡선의 오른쪽 절반이고(소문항 (ii)에서 $k \approx 1.60$), $x = 0$에서 0으로부터 k까지 뛰어오르는 불연속이 있다.



소문항 (ii)

전체 확률은 1이어야 한다. g는 0에 대하여 대칭이고 실수 전체에서 적분하면 1이므로, $[0, \infty)$에서의 적분은 $\frac{1}{2}$이다.

$$\int_0^\infty \frac{2}{\sqrt{2\pi}} \mathrm{e}^{-2x^2}\,\mathrm{d}x = \frac{1}{2} \quad \Longrightarrow \quad \int_0^\infty \mathrm{e}^{-2x^2}\,\mathrm{d}x = \frac{\sqrt{2\pi}}{4}.$$

따라서 $k \cdot \dfrac{\sqrt{2\pi}}{4} = 1$이고, 이로부터

$$k = \frac{4}{\sqrt{2\pi}} = 2\sqrt{\frac{2}{\pi}} \approx 1.596.$$

을 얻는다. $x \geq 0$에서 $f(x) = 2g(x)$임에 주목하자. 이것은 X가 $|Y|$와 같은 분포를 가진다는 것을 확인해 준다.

소문항 (iii)

평균(mean)의 경우, 인수 x 덕분에 적분이 초등적으로 계산된다. $x\,\mathrm{e}^{-2x^2}$가 e^{-2x^2}의 도함수의 상수배이기 때문이다.

$$\mathrm{E}(X) = k\int_0^\infty x\,\mathrm{e}^{-2x^2}\,\mathrm{d}x = k\Big[-\tfrac{1}{4}\mathrm{e}^{-2x^2}\Big]_0^\infty = \frac{k}{4} = \frac{1}{\sqrt{2\pi}} \approx 0.399.$$

? **생각해 보기** 긴 부분적분(integration by parts) 없이 $\displaystyle\int_0^\infty x^2\,\mathrm{e}^{-2x^2}\,\mathrm{d}x$를 어떻게 구할 수 있을까?

다시 정규분포를 이용한다. 함수 $x^2 g(x)$는 우함수이므로

$$\mathrm{E}(X^2) = \int_0^\infty x^2 \cdot 2g(x)\,\mathrm{d}x = \int_{-\infty}^\infty x^2 g(x)\,\mathrm{d}x = \mathrm{E}(Y^2) = \mathrm{Var}(Y) = \frac{1}{4},$$

이다. $\mathrm{E}(Y) = 0$이기 때문이다. ($u = x$, $\mathrm{d}v = x\mathrm{e}^{-2x^2}\,\mathrm{d}x$로 놓고 부분적분을 해도 같은 결과 $\int_0^\infty x^2\mathrm{e}^{-2x^2}\,\mathrm{d}x = \frac{1}{4}\int_0^\infty \mathrm{e}^{-2x^2}\,\mathrm{d}x = \frac{1}{4k}$를 얻는다.) 따라서 분산(variance)은

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{1}{4} - \frac{1}{2\pi} = \frac{\pi - 2}{4\pi} \approx 0.0908.$$

소문항 (iv)

중앙값 m은 $\mathrm{P}(X \le m) = \frac{1}{2}$를 만족한다. X는 $|Y|$와, Y는 $\frac{1}{2}Z$와 같은 분포를 가지므로, 표준정규분포(standard normal distribution)의 대칭성을 이용하면

$$\mathrm{P}(X \le m) = \mathrm{P}(|Z| \le 2m) = 2\Phi(2m) - 1,$$

이다. (직접 계산하면 $\mathrm{P}(X \le m) = \int_0^m 2g(x)\,\mathrm{d}x = 2(\mathrm{P}(Y \le m) - \frac{1}{2})$이다.) 이것을 $\frac{1}{2}$과 같게 놓으면

$$\Phi(2m) = \frac{3}{4}.$$

을 얻는다. 표에서 $\Phi^{-1}(0.75) = 0.6745$이므로 $2m = 0.6745$이고

$$m \approx 0.337 \quad (\text{유효숫자 } 3\text{자리}).$$

검산

$k = 2\sqrt{2/\pi}$로 두고 f를 $[0, 10]$에서 수치적으로(심프슨 공식(Simpson's rule)으로) 적분하면 전체 확률 1.0000, $\mathrm{E}(X) = 0.39894$, $\mathrm{Var}(X) = 0.090845$를 얻으며, 이는 $1/\sqrt{2\pi}$ 및 $(\pi - 2)/(4\pi)$와 일치한다. $\int_0^m f(x)\,\mathrm{d}x = \frac{1}{2}$를 수치적으로 풀면 $m = 0.33724$이다. 표준편차가 $\frac{1}{2}$인 정규분포 값 200 000개로 $|Y|$를 모의실험하면 평균 0.399, 분산 0.091을 얻었고, 값의 50.0%가 0.337보다 작았다. 상식 점검: 중앙값 0.337이 평균 0.399보다 작은데, 오른쪽 꼬리가 긴 분포라면 마땅히 그래야 한다.

흔한 실수

- 지수로부터 $\sigma^2 = 2$ 또는 $\sigma = \frac{1}{4}$로 잘못 읽는 것: $2x^2$를 $x^2/(2\sigma^2)$와 맞추면 $\sigma^2 = \frac{1}{4}$가 된다.
- 인수 2를 빠뜨리는 것: 실수 직선의 절반에서는 정규분포의 밀도를 적분해도 $\frac{1}{2}$밖에 되지 않는다.
- $\Phi(2m) = \frac{1}{2}$를 풀어 $m = 0$을 얻는 것. m 아래에 있는 X의 확률 절반은 Y의 확률 중 가운데 절반에 해당하므로, 조건은 $\Phi(2m) = \frac{3}{4}$이다.

STEP 너머

X는 **반정규**(*half-normal*) 확률변수이며, Y가 정규분포를 따를 때의 $|Y|$는 **확률변수의 변환**(*transformation of a random variable*)의 가장 간단한 예이다. 일반적으로 $X = h(Y)$이면, X의 밀도는 모든 원상(preimage)에서의 Y의 밀도를 더하여 얻으며, 여기서는 그것이 인수 2이다. 같은 아이디어로 Z^2가 자유도(degree of freedom) 1인 카이제곱분포(chi-squared distribution)를 따른다는 것을 보일 수 있고, 이것은 통계학의 많은 부분의 출발점이 된다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 확률변수의 함수를 꼼꼼하게 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.