풀이

STEP III 2012, 문제 13

확률과 통계 • 절단정규분포

> **STEP III 2012, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/12-s3-q13

핵심 아이디어

구간 위에서의 조건부 기댓값(conditional expectation)은 그 구간에서 $z\,f(z)$를 적분한 값을 그 구간의 확률로 나눈 것이다. 표준정규분포(standard Normal)에서 $z\,\mathrm{e}^{-z^2/2}$는 정확히 $-\mathrm{e}^{-z^2/2}$의 도함수이므로, 이 적분은 정확하게 계산할 수 있다. 그다음에 할 일은 $X = \mu + \sigma Z$로 표준화(standardising)하는 것과, X의 부호에 따라 $|X|$를 나누어 다루는 것뿐이다.

풀이

이 풀이 전체에서 Z의 확률밀도함수(probability density function)를 $\phi(z) = \dfrac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-z^2/2}$로 쓴다. 그러면 $\Phi'(z) = \phi(z)$이고 $\phi'(z) = -z\,\phi(z)$이다.

소문항 (i)

$a < Z < b$일 때 Z의 조건부 밀도(conditional density)는 (a, b)에서 $\phi(z)/\mathrm{P}(a < Z < b)$이고, $\mathrm{P}(a < Z < b) = \Phi(b) - \Phi(a)$이다. 따라서

$$\begin{aligned}
\mathrm{E}(Z \mid a < Z < b) &= \frac{1}{\Phi(b) - \Phi(a)} \int_a^b z\,\phi(z)\,\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}\,(\Phi(b) - \Phi(a))} \Big[-\mathrm{e}^{-z^2/2} \Big]_a^b \\
&= \frac{\mathrm{e}^{-a^2/2} - \mathrm{e}^{-b^2/2}}{\sqrt{2\pi}\,(\Phi(b) - \Phi(a))},
\end{aligned}$$

이다. 여기서 $\dfrac{\mathrm{d}}{\mathrm{d}z}(-\mathrm{e}^{-z^2/2}) = z\,\mathrm{e}^{-z^2/2}$임을 이용하였다. 이것이 구하는 결과이다. 적분이 수렴(converge)하므로, $\mathrm{e}^{-\infty} = 0$, $\Phi(\infty) = 1$로 읽으면 이 결과는 $b = \infty$일 때(또는 $a = -\infty$일 때)에도 성립한다.

소문항 (ii)

$Z \sim \mathrm{N}(0,1)$, $\sigma > 0$으로 두고 $X = \mu + \sigma Z$라 쓰자. $\sigma > 0$이므로

$$X > 0 \iff Z > -\mu/\sigma,$$

이고, 따라서 $X > 0$이라는 조건을 거는 것은 $Z > -\mu/\sigma$라는 조건을 거는 것과 같다. 기댓값은 조건부로 다룰 때에도 선형(linear)이므로

$$\mathrm{E}(X \mid X > 0) = \mathrm{E}(\mu + \sigma Z \mid Z > -\mu/\sigma) = \mu + \sigma\,\mathrm{E}(Z \mid Z > -\mu/\sigma),$$

이고, 이것이 구하는 결과이다.

$|X|$의 평균. $c = \mu/\sigma$로 두자. 소문항 (i)에서 $a = -c$, $b = \infty$로 놓으면

$$\mathrm{E}(Z \mid Z > -c) = \frac{\mathrm{e}^{-c^2/2}}{\sqrt{2\pi}\,(1 - \Phi(-c))} = \frac{\phi(c)}{p}, \qquad \text{여기서} p = \mathrm{P}(X > 0) = 1 - \Phi(-c).$$

을 얻는다. 앞의 식에 p를 곱하면

$$\mathrm{E}(X;\, X > 0) := \mathrm{E}(X \mid X > 0)\,\mathrm{P}(X > 0) = \mu p + \sigma\phi(c).$$

이다.

? **생각해 보기** 두 번째 적분 없이 $X < 0$ 부분의 기여를 어떻게 구할 수 있을까?

두 부분을 더하면 전체 평균이 된다. 즉 $\mathrm{E}(X;\, X > 0) + \mathrm{E}(X;\, X < 0) = \mathrm{E}(X) = \mu$이다($X = 0$인 사건의 확률은 0이다). 따라서 $\mathrm{E}(X;\, X < 0) = \mu - \mu p - \sigma\phi(c)$이다. $X < 0$에서는 $|X| = -X$이므로

$$\begin{aligned} m = \mathrm{E}|X| &= \mathrm{E}(X;\, X > 0) - \mathrm{E}(X;\, X < 0) \\ &= 2(\mu p + \sigma\phi(c)) - \mu \\ &= \mu(2p - 1) + 2\sigma\phi(c). \end{aligned}$$

이다. 마지막으로 $2p - 1 = 1 - 2\Phi(-\mu/\sigma)$이고 $2\phi(c) = \sqrt{2/\pi}\,\mathrm{e}^{-\mu^2/(2\sigma^2)}$이므로

$$m = \mu(1 - 2\Phi(-\mu/\sigma)) + \sigma\sqrt{2/\pi}\,\exp(-\tfrac{1}{2}\mu^2/\sigma^2),$$

을 얻고, 이것이 구하는 결과이다.

$|X|$의 분산. 핵심은 $|X|^2 = X^2$이라는 점이다. 따라서 이차 모멘트(second moment)는 새로 계산할 필요가 없다.

$$\mathrm{E}(|X|^2) = \mathrm{E}(X^2) = \mathrm{Var}(X) + (\mathrm{E}X)^2 = \sigma^2 + \mu^2.$$

그러므로

$$\mathrm{Var}|X| = \mu^2 + \sigma^2 - m^2.$$

이다.

검산

- $\mu = 0$이면 $|X|$는 반정규분포(half-Normal)를 따르고, 공식은 잘 알려진 값 $m = \sigma\sqrt{2/\pi}$를 준다. 이때 분산은 $\sigma^2(1 - 2/\pi)$로 양수이다.
- $\mu \gg \sigma$이면 X는 거의 항상 양수이고, $\Phi(-\mu/\sigma) \approx 0$이며 지수 항은 매우 작다. 따라서 $m \approx \mu$, $\mathrm{Var}|X| \approx \sigma^2$이 되는데, 마땅히 그래야 한다.
- m에 대한 공식은 μ를 $-\mu$로 바꾸어도 변하지 않는다($1 - 2\Phi(-c) = -(1 - 2\Phi(c))$를 이용한다). 이는 대칭성 $|X| \sim |-X|$와 맞아떨어진다.
- 수치적으로 $\mu = 1$, $\sigma = 2$이면 $m \approx 1.7912$, $\mathrm{Var}|X| \approx 1.7917$을 얻는데, 이는 직접 수치적분한 값 및 2×10^5개의 값을 이용한 시뮬레이션 결과와 일치한다.

흔한 실수

- 소문항 (i)에서 $\Phi(b) - \Phi(a)$로 나누는 것을 잊는 실수. 적분만으로는 조건부 기댓값이 아니라 $\mathrm{E}(Z;\, a < Z < b)$가 된다.
- $\mathrm{E}|X| = \mathrm{E}(X \mid X > 0)$이라고 쓰는 실수. 조건부 기댓값에는 각각 $\mathrm{P}(X > 0)$과 $\mathrm{P}(X < 0)$의 가중치를 주어야 한다.
- $|X|^2 = X^2$을 이용하면 곧바로 얻을 수 있는데도 $\mathrm{E}(X^2)$를 또 다른 적분으로 계산하는 것.

STEP 너머

$|X|$의 분포는 **접힌 정규분포**(folded Normal)이고, 소문항 (i)의 조건부 분포는 **절단 정규분포**(truncated Normal)이다. 비율 $\phi(c)/\big(1 - \Phi(-c)\big)$는 역밀스 비(inverse Mills ratio)로, 데이터가 중도절단(censored)되거나 선택(selected)된 계량경제학 모형에서 중심적인 역할을 한다. 요령 $\phi'(z) = -z\phi(z)$는 정규근사를 위한 스타인 방법(Stein's method)의 출발점이기도 하다. 조건부 기댓값에 대해 더 알고 싶다면 Blitzstein and Hwang, *Introduction to Probability*(확률론 입문)를 참고하라.

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.