풀이

STEP II 2004, 문제 12

확률과 통계 • 연속 분포

> **STEP II 2004, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/04-s2-q12

핵심 아이디어

모든 것은 하나의 역도함수(antiderivative)에서 나온다. xe^{-ax^2}는 상수배를 제외하면 e^{-ax^2}의 도함수이다. 이로부터 정규화 상수(normalising constant)와 누적분포함수(cumulative distribution function)의 닫힌 꼴을 얻으므로, 중앙값(median)은 지수방정식 하나를 풀어서 구할 수 있다. 최빈값(mode)은 곡선이 최고점에 이르는 곳이다. 단, 그 최고점이 $[0,1]$ 안에 있어야 한다. 그렇지 않으면 밀도(density)는 $x=1$에서도 여전히 증가하고 있다. 그러면 조건부 확률(conditional probability)은 m과 h에서의 누적분포함수 값만 있으면 구할 수 있다.

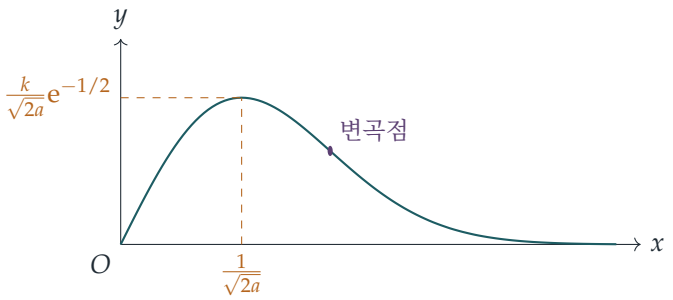
풀이

그래프 그리기

$x \geq 0$에 대하여 $g(x) = kxe^{-ax^2}$라 하자. 그러면 $g(0)=0$이고, $x>0$이면 $g(x)>0$이며, 곱의 미분법(product rule)에 의해

$$g'(x) = ke^{-ax^2}(1-2ax^2),$$

이다. 이 값은 $0 \leq x < 1/\sqrt{2a}$에서 양수이고 그 이후로는 음수이다. 따라서 g는 원점에서 (그곳에서의 기울기는 k) 증가하여 $x = 1/\sqrt{2a}$에서 유일한 극댓값 $y = \dfrac{k}{\sqrt{2a}}e^{-1/2}$에 이른 뒤 감소한다. $x \to \infty$일 때 지수함수가 일차 인수를 압도하므로 $y \to 0$이다. 즉 x축이 점근선(asymptote)이다. (한 번 더 미분하면 $x = \sqrt{3/(2a)}$에서 변곡점(point of inflection)이 있음을 알 수 있고, 그곳에서 곡선은 위로 볼록(concave)에서 아래로 볼록(convex)으로 바뀐다.)



상수 k

$\frac{d}{dx}e^{-ax^2} = -2axe^{-ax^2}$이므로, $0 \le t \le 1$에 대하여

$$\int_0^t kxe^{-ax^2}\,dx = \left[-\frac{k}{2a}e^{-ax^2}\right]_0^t = \frac{k}{2a}(1-e^{-at^2}).$$

전체 확률은 1이므로 $t = 1$로 놓으면 $\frac{k}{2a}(1-e^{-a}) = 1$, 즉

$$k = \frac{2a}{1-e^{-a}},$$

이고, 이것이 구하려던 결과이다. ($a > 0$이므로 분모는 양수이다.) 이것을 다시 대입하면 $[0,1]$에서의 누적분포함수는

$$F(t) = P(X \le t) = \frac{1-e^{-at^2}}{1-e^{-a}}.$$

최빈값

밀도함수는 앞에서 그린 곡선을 $x = 1$에서 잘라 낸 것이다. 그래프에서 알 수 있듯이, $[0,1]$에서 밀도는 극대점이 구간 안에 있으면 그 극대점에서, 그렇지 않으면 오른쪽 끝점에서 가장 크다.

- $a > \frac{1}{2}$이면 $2a > 1$이므로 $1/\sqrt{2a} < 1$이다. 최고점이 구간 안에 있으므로 $m = \frac{1}{\sqrt{2a}}$이다.
- $a < \frac{1}{2}$이면 $1/\sqrt{2a} > 1$이므로 $[0,1)$ 전체에서 $f'(x) > 0$이다. 밀도는 구간 전체에서 증가하므로 $m = 1$이다.

($a = \frac{1}{2}$일 때는 두 공식 모두 $m = 1$을 준다.)

중앙값

중앙값 h는 $F(h) = \frac{1}{2}$를 만족한다.

$$1 - \mathrm{e}^{-ah^2} = \tfrac{1}{2}\left(1 - \mathrm{e}^{-a}\right) \quad \Longleftrightarrow \quad \mathrm{e}^{-ah^2} = \frac{1 + \mathrm{e}^{-a}}{2}.$$

로그를 취하고 $h > 0$을 이용하면

$$h = \sqrt{\frac{1}{a} \ln\left(\frac{2}{1 + \mathrm{e}^{-a}}\right)}.$$

$1 + \mathrm{e}^{-a} < 2$이므로 로그 값이 양수이고, 따라서 이 값은 실수이다.

$h > m$은 언제 성립하는가?

$a \leq \frac{1}{2}$이면 $m = 1$이고 $h < 1$이므로 (확률의 절반이 h 아래에 있고, 1 위에는 확률이 전혀 없다) $h < m$이다. 따라서 $h > m$은 $a > \frac{1}{2}$일 때만 가능하고, 이때 $m = 1/\sqrt{2a}$이다. 둘 다 양수이므로 제곱을 비교하면

$$\begin{aligned} h > m &\iff \frac{1}{a} \ln\left(\frac{2}{1 + \mathrm{e}^{-a}}\right) > \frac{1}{2a} \iff \ln\left(\frac{2}{1 + \mathrm{e}^{-a}}\right) > \frac{1}{2} \\ &\iff \frac{2}{1 + \mathrm{e}^{-a}} > \mathrm{e}^{1/2} \iff 1 + \mathrm{e}^{-a} < 2\mathrm{e}^{-1/2} \\ &\iff \mathrm{e}^{-a} < 2\mathrm{e}^{-1/2} - 1 \iff a > -\ln\left(2\mathrm{e}^{-1/2} - 1\right). \end{aligned}$$

첫 번째 단계에서는 $a > 0$을 곱했고, 세 번째 단계에서는 exp가 증가함수라는 사실을, 마지막 단계에서는 ln이 증가함수라는 사실을 이용했다. 마지막 단계에는 $2\mathrm{e}^{-1/2} - 1 > 0$도 필요한데, 이는 $\mathrm{e}^{1/2} < 2$이므로 (왜냐하면 $\mathrm{e} < 4$이므로) 성립한다.

> **생각해 보기** 위의 동치 사슬은 $a > \frac{1}{2}$를 가정했다. 그렇다면 결과에 이것을 추가 조건으로 붙여야 하지 않을까?

그렇지 않다. 바로 그 때문에 다음 소문항에서 그것을 묻는 것이다. 아래에서 $c = -\ln\left(2\mathrm{e}^{-1/2} - 1\right)$가 $\frac{1}{2}$보다 크다는 것을 보인다. 따라서 $a > c$이면 자동으로 $a > \frac{1}{2}$이고, 공식 $m = 1/\sqrt{2a}$가 적용되며, 위의 사슬에 의해 $h > m$이다. $a \leq \frac{1}{2}$인 경우와 합하면, $h > m$일 필요충분조건은 $a > c$이다.

상수가 $\frac{1}{2}$보다 크다

ln은 증가함수이므로

$$-\ln\left(2\mathrm{e}^{-1/2} - 1\right) > \tfrac{1}{2} \iff 2\mathrm{e}^{-1/2} - 1 < \mathrm{e}^{-1/2} \iff \mathrm{e}^{-1/2} < 1,$$

이고, 마지막 부등식은 참이다. (수치로는 $c \approx 1.546$이다.)

조건부 확률

$a > c$라 하자. 그러면 $m < h$이므로 사건 $\{X > m\} \cap \{X < h\}$는 사건 $\{m < X < h\}$와 같다. $P(X < h) = \frac{1}{2}$이므로

$$P(X > m \mid X < h) = \frac{P(m < X < h)}{P(X < h)} = \frac{F(h) - F(m)}{\frac{1}{2}} = 1 - 2F(m).$$

$m = 1/\sqrt{2a}$이면 $am^2 = \frac{1}{2}$이므로 $F(m) = \dfrac{1 - \mathrm{e}^{-1/2}}{1 - \mathrm{e}^{-a}}$이다. 따라서

$$P(X > m \mid X < h) = \frac{(1 - \mathrm{e}^{-a}) - 2(1 - \mathrm{e}^{-1/2})}{1 - \mathrm{e}^{-a}} = \frac{2\mathrm{e}^{-1/2} - \mathrm{e}^{-a} - 1}{1 - \mathrm{e}^{-a}},$$

이고, 이것이 구하려던 결과이다.

검산

- $a = c$에서 분자 $2\mathrm{e}^{-1/2} - \mathrm{e}^{-a} - 1$은 정확히 0이며, 이는 그곳에서 $m = h$인 것과 맞아떨어진다. $a > c$이면 분자는 양수로, 확률이라면 마땅히 그래야 한다.
- $a \to \infty$일 때 답은 $2\mathrm{e}^{-1/2} - 1 \approx 0.213$으로 수렴(converge)하며, 이는 1보다 작다.
- $a = 2.5$이면 $k \approx 5.447$, $m \approx 0.447$, $h \approx 0.496$이고, 밀도를 수치적분하면 $P(X > m \mid X < h) \approx 0.1427$을 얻어 공식과 일치한다. $a = 1$이면 $m \approx 0.707 > h \approx 0.616$이며, 이는 $1 < c$와 부합한다.

흔한 실수

- 모든 경우에 $m = 1/\sqrt{2a}$라고 답하는 것. $a < \frac{1}{2}$이면 극대점이 밀도가 0인 $x = 1$ 너머에 있다.
- F 안의 k가 정규화된 값이어서 $F(1) = 1$이 되어야 한다는 것을 잊는 것.
- $P(X > m \mid X < h) = P(X > m)/P(X < h)$라고 쓰는 것. 분자는 두 사건이 함께 일어날 확률이어야 한다.

STEP 너머

이 밀도는 절단된 레일리 분포(truncated Rayleigh distribution)이다. $x = 1$에서 잘라 내지 않으면, $k = 2a$인 $kx\mathrm{e}^{-ax^2}$는 두 좌표가 서로 독립인 정규분포(normal distribution) 확률변수인 점과 원점 사이의 거리의 분포이다. 평균, 중앙값, 최빈값을 비교하고 절단이 이들을 어떻게 이동시키는지 살펴보는 것은 처음 배우는 통계학 강좌에서 거듭 등장하는 주제이

다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 연속확률변수와 그 밀도함수, 분포함수에 대해 더 읽어 보기에 좋은 책이다.

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.