풀이

STEP II 2007, 문제 14

확률과 통계 • 구간별로 정의된 확률밀도함수

> **STEP II 2007, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/07-s2-q14

핵심 아이디어

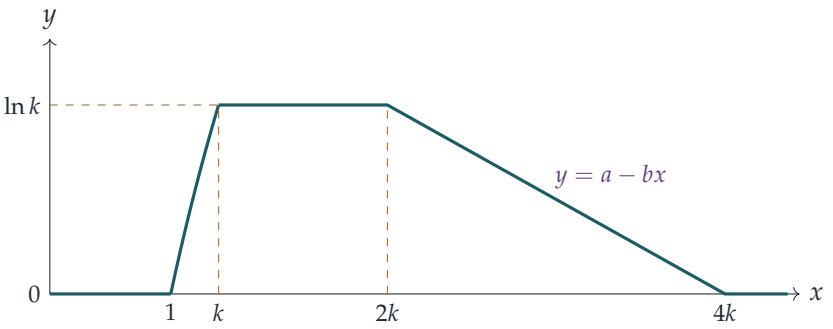
여기서는 **연속**(*continuous*)이라는 단어가 큰 역할을 한다. 구간별로 정의된 확률밀도함수(probability density function)는 모든 경계점에서 이어져야 하고, 이로부터 곧바로 a와 b에 대한 두 방정식을 얻는다. 남은 하나의 조건, 즉 그래프 아래의 전체 넓이가 1이라는 조건이 k를 결정한다. 중앙값(median)을 구할 때는 넓이를 구간별로 더해 나가다가 절반을 넘는 곳을 찾는다.

풀이

전체에 걸쳐 $k > 1$이다. 그렇지 않으면 구간 $1 \le x \le k$가 공집합이 되고, 어쨌든 밀도가 음이 아니면서 $x = 1$ 이후에서 항상 0이 되지 않으려면 $\ln k > 0$이어야 한다.

소문항 (i)

$[1, k]$에서 그래프는 익숙한 로그 곡선으로, $x = 1$일 때 0에서 시작하며(따라서 값이 0인 구간과 연속적으로 이어진다) 증가하고 위로 볼록(concave)하다. $x = k$에서 높이 $\ln k$에 도달하고, $x = 2k$까지 $\ln k$로 일정하게 유지되다가, $x = 4k$에서 0이 되도록 직선으로 내려가며(연속성이 바로 이렇게 되도록 강제한다는 것을 아래에서 보인다), 그 이후로는 0이다. 그림에는 소문항 (ii)에서 구하는 값 $k = \mathrm{e}^{1/3} \approx 1.40$을 사용했으므로, 평평한 부분의 높이는 $\frac{1}{3}$이다.



소문항 (ii)

연속성. $x = k$에서는 두 식 모두 $\ln k$를 주므로 새로운 정보가 없다. $x = 2k$와 $x = 4k$에서 연속이려면

$$a - 2bk = \ln k, \qquad a - 4bk = 0.$$

이어야 한다. 두 식을 빼면 $2bk = \ln k$이므로

$$b = \frac{\ln k}{2k}, \qquad a = 4bk = 2\ln k.$$

이다.

전체 확률. 넓이를 세 부분으로 나눈다. 부분적분(integration by parts)을 하면($u = \ln x$, $\mathrm{d}v = \mathrm{d}x$로 놓는다)

$$\int_1^k \ln x \,\mathrm{d}x = \Big[x\ln x - x\Big]_1^k = k\ln k - k + 1.$$

이다. 평평한 부분은 너비 k, 높이 $\ln k$인 직사각형이므로 넓이는 $k\ln k$이다. 마지막 부분은 밑변 $4k - 2k = 2k$, 높이 $\ln k$인 삼각형이므로 넓이는 $\frac{1}{2}\cdot 2k\cdot \ln k = k\ln k$이다. 따라서

$$1 = (k\ln k - k + 1) + k\ln k + k\ln k = 3k\ln k - k + 1,$$

이고, 이를 정리하면

$$k(3\ln k - 1) = 0.$$

이다. $k > 1$이므로 $k = 0$일 수 없고, 따라서 $\ln k = \frac{1}{3}$이며

$$k = \mathrm{e}^{1/3} \approx 1.396, \qquad a = 2\ln k = \tfrac{2}{3}, \qquad b = \frac{\ln k}{2k} = \frac{1}{6\,\mathrm{e}^{1/3}} \approx 0.119.$$

이다. $b > 0$이므로 마지막 부분은 그림처럼 실제로 내려가는 직선임에 주목하자.

소문항 (iii)

확률을 누적해 가며 중앙값 m이 어느 구간에 있는지 찾는다. $\ln k = \frac{1}{3}$을 이용하면

$$\mathrm{P}(X \le k) = k\ln k - k + 1 = 1 - \tfrac{2k}{3} \approx 0.070,$$

이고, 평평한 부분이 $k\ln k = \frac{k}{3}$를 더하므로

$$\mathrm{P}(X \le 2k) = 1 - \tfrac{2k}{3} + \tfrac{k}{3} = 1 - \tfrac{k}{3} \approx 0.535.$$

이다.

> **생각해 보기** 중앙값은 그래프의 어느 부분에 있는가?

$0.070 < \frac{1}{2} < 0.535$이므로 중앙값은 평평한 부분, 즉 $k \le m \le 2k$에 있다. 이 구간에서 밀도는 $\frac{1}{3}$으로 일정하므로

$$\mathrm{P}(X \le m) = 1 - \tfrac{2k}{3} + \tfrac{1}{3}(m - k) = \tfrac{1}{2}.$$

이다. 양변에 3을 곱하면 $3 - 2k + m - k = \frac{3}{2}$이므로

$$m = 3k - \tfrac{3}{2} = 3\mathrm{e}^{1/3} - \tfrac{3}{2} \approx 2.69.$$

이다. 이 값은 예상대로 $k \approx 1.40$과 $2k \approx 2.79$ 사이에 있다.

검산

- $k = \mathrm{e}^{1/3}$, $a = \frac{2}{3}$, $b = 1/(6\mathrm{e}^{1/3})$일 때 밀도를 수치적분(중점 규칙)하면 전체 넓이는 1.0000000이고, $\int_1^{2.68684} f(x)\,\mathrm{d}x = 0.5000000$이다.
- 중앙값에 대한 상식적 검토: 확률의 절반을 조금 넘는 양(약 0.535)이 $2k$의 왼쪽에 있으므로, 중앙값은 $2k \approx 2.79$보다 약간 왼쪽에 있어야 한다. 실제로 $m \approx 2.69$이다.

흔한 실수

- $\int_1^k \ln x\,\mathrm{d}x$에서 아래끝을 빠뜨리는 것. $x = 1$에서 $-[x\ln x - x]$가 주는 $+1$ 덕분에 k에 대한 방정식이 깔끔하게 인수분해된다.
- 중앙값이 받침(support) $[1, 4k]$의 "가운데"에 있으니 기울어진 부분에 있을 것이라고 가정하는 것. 항상 넓이를 먼저 누적해 보아야 한다.

STEP 너머

여기서는 밀도를 여러 조각으로 만들고, 상수들을 연속성과 정규화(normalisation) 조건으로 결정했다. 대학에서는 **누적분포함수**(*cumulative distribution function*) $F(x) = \mathrm{P}(X \le x)$를 기본 대상으로 다루게 된다. 누적분포함수는 이산이든 연속이든 모든 확률변수에 대해 존재하며, 중앙값 같은 양은 단순히 $F(m) = \frac{1}{2}$의 해이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)이 이러한 관점을 명료하게 전개한다.

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.