풀이

STEP III 1991, 문제 16

확률과 통계 • 결합 분포와 주변 분포

STEP III 1991, 16번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/91-s3-q16

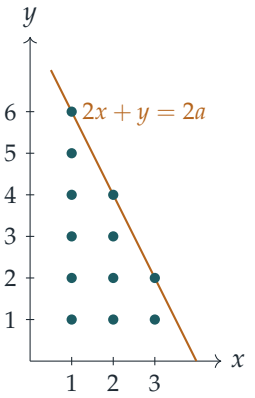
핵심 아이디어

여기서 다루는 것은 모두 삼각형 모양으로 놓인 격자점(lattice points) 위에서의 유한합이다. 먼저 영역을 그려 보자. 각 x에 대해 어떤 y가 허용되는지, 그리고 각 y에 대해 어떤 x가 허용되는지 알아야 한다. 범위를 올바르게 잡고 나면 각 주변분포(marginal distribution)는 등차수열(arithmetic progression)의 합이 된다. 유일하게 미묘한 점은 주어진 y에 대해 허용되는 가장 큰 x가 $\lfloor (2a-y)/2 \rfloor$이고, 이 값이 y가 짝수인지 홀수인지에 따라 달라진다는 것이다.

풀이

영역

$x \geq 1, y \geq 1, 2x + y \leq 2a$가 성립해야 한다. $y \geq 1$이므로 $2x \leq 2a - 1$이어야 하고, 따라서 $x \leq a - 1$이다. $1 \leq x \leq a - 1$인 고정된 x에 대해 허용되는 값은 $y = 1, 2, \ldots, 2a - 2x$이다. 마찬가지로 $x \geq 1$이므로 $y \leq 2a - 2$이다.



그림은 $a = 4$인 경우를 보여 준다. 열 x에는 $2a - 2x$개의 점이 있다.

X의 주변분포와 c의 값

x를 고정하고 $n = 2a - 2x$로 쓰자. 등차수열 $2x + 1, \ldots, 2x + n$을 더하면

$$\begin{aligned} \mathrm{P}(X = x) &= c \sum_{y=1}^{n} (2x + y) = c \left(2xn + \frac{n(n+1)}{2} \right) = c \, \frac{n}{2} \, (4x + n + 1) \\ &= c \, (a - x)(2a + 2x + 1), \qquad x = 1, 2, \ldots, a - 1, \end{aligned}$$

이고, 그 밖의 경우에는 $\mathrm{P}(X = x) = 0$이다.

확률을 모두 더하면 1이 되어야 한다. $k = a - x$로 놓으면 이 값은 1부터 $a - 1$까지 변하고, $2a + 2x + 1 = 4a + 1 - 2k$이며

$$\begin{aligned} \sum_{k=1}^{a-1} k(4a + 1 - 2k) &= (4a + 1) \frac{(a-1)a}{2} - 2 \cdot \frac{(a-1)a(2a-1)}{6} \\ &= a(a-1) \left(\frac{4a+1}{2} - \frac{2a-1}{3} \right) = \frac{a(a-1)(8a+5)}{6}. \end{aligned}$$

이다. 이 값에 c를 곱한 것을 1로 놓으면

$$c = \frac{6}{a(a-1)(8a+5)},$$

이고, 이것이 요구된 결과이다.

Y의 주변분포

$1 \le y \le 2a - 2$인 y를 고정하자. 허용되는 x는 $1, 2, \ldots, N$이고, 여기서 N은 $2N \le 2a - y$를 만족하는 가장 큰 정수이다. 이러한 N개의 항에 대해

$$\sum_{x=1}^{N} (2x + y) = N(N+1) + yN = N(N + 1 + y).$$

y가 짝수인 경우. 이때 $2a-y$는 짝수이고 $N=\frac{1}{2}(2a-y)$이므로 $N+1+y=\frac{1}{2}(2a+2+y)$이고

$$\mathrm{P}(Y=y)=c\cdot\frac{(2a-y)(2a+2+y)}{4}=\frac{3(2a-y)(2a+2+y)}{2a(a-1)(8a+5)},$$

이다. 여기서 c의 값을 이용하였다. 이것이 요구된 결과이다.

? **생각해 보기** y가 홀수일 때는 왜 $N=\frac{1}{2}(2a-y)$를 쓸 수 없을까?

y가 홀수이면 $\frac{1}{2}(2a-y)$는 정수가 아니다. $2x\le 2a-y$를 만족하는 가장 큰 정수 x는 $N=\frac{1}{2}(2a-y-1)$이다. 내림을 하면 절반이 떨어져 나가기 때문이다.

y가 홀수인 경우. $N=\frac{1}{2}(2a-1-y)$로 놓으면 $N+1+y=\frac{1}{2}(2a+1+y)$이므로

$$\mathrm{P}(Y=y)=c\cdot\frac{(2a-1-y)(2a+1+y)}{4}=\frac{3(2a-1-y)(2a+1+y)}{2a(a-1)(8a+5)}.$$

(두 경우는 같은 분모 위에 각각 $3\big((2a+1)^2-(y+1)^2\big)$과 $3\big(4a^2-(y+1)^2\big)$로 쓸 수 있다.)

Y의 기댓값(expectation)

$y\,\mathrm{P}(Y=y)$를 더하려면 짝수인 y와 홀수인 y를 따로 다루어야 한다. 결합분포(joint distribution)로 돌아가서 각 열을 따라 더하는 편이 더 깔끔하다. 고정된 x에 대해, 앞에서처럼 $n=2a-2x$로 두면

$$\sum_{y=1}^{n}y(2x+y)=2x\cdot\frac{n(n+1)}{2}+\frac{n(n+1)(2n+1)}{6}=\frac{n(n+1)}{6}\big(6x+2n+1\big).$$

다시 $k=a-x$로 놓으면 $n=2k$이고 $6x+2n+1=6a-2k+1$이다. 열의 합은 다음과 같이 된다.

$$\frac{2k(2k+1)(6a+1-2k)}{6}=\frac{k(2k+1)(6a+1-2k)}{3}.$$

전개하면 $k(2k+1)(6a+1-2k)=(6a+1)(2k^2+k)-4k^3-2k^2$이다. $a-1$까지의 표준적인 합 공식을 이용하면

$$\sum_{k=1}^{a-1}(2k^2+k)=\frac{a(a-1)(2a-1)}{3}+\frac{a(a-1)}{2}=\frac{a(a-1)(4a+1)}{6},$$

$$\sum_{k=1}^{a-1}k^3=\frac{a^2(a-1)^2}{4},\qquad\sum_{k=1}^{a-1}k^2=\frac{a(a-1)(2a-1)}{6}.$$

따라서

$$\begin{aligned}&\sum_{k=1}^{a-1} k(2k+1)(6a+1-2k)\\&\quad= a(a-1)\left(\frac{(6a+1)(4a+1)}{6} - a(a-1) - \frac{2a-1}{3}\right)\\&\quad= \frac{a(a-1)}{6}\left(24a^2+10a+1-6a^2+6a-4a+2\right)\\&\quad= \frac{a(a-1)(6a^2+4a+1)}{2}.\end{aligned}$$

그러므로

$$\begin{aligned}\mathrm{E}(Y) &= \frac{c}{3}\cdot\frac{a(a-1)(6a^2+4a+1)}{2} = \frac{6}{a(a-1)(8a+5)}\cdot\frac{a(a-1)(6a^2+4a+1)}{6}\\&= \frac{6a^2+4a+1}{8a+5}.\end{aligned}$$

검산

- $a=2$이면 점은 $(1,1)$과 $(1,2)$ 두 개뿐이고 가중치는 각각 $3c$와 $4c$이므로 $c=\frac{1}{7}$이다. 이는 $6/(2\cdot1\cdot21)$과 일치한다. 그러면 $\mathrm{E}(Y)=\frac{3+8}{7}=\frac{11}{7}$이고, 공식도 $\frac{33}{21}=\frac{11}{7}$을 준다.
- X와 Y의 주변분포 공식(두 홀짝성(parity) 모두)과 $\mathrm{E}(Y)$를 2부터 29까지의 모든 a에 대해 분수를 이용한 정확한 전수 계산으로 확인하였다. 예를 들어 $a=3$이면 $\mathrm{E}(Y)=\frac{67}{29}$이다.
- a가 크면 $\mathrm{E}(Y)\approx\frac{3}{4}a$이다. 가중치 $2x+y$는 비스듬한 변 근처의 점에 더 큰 비중을 주므로, Y는 삼각형 위에 균등한 가중치를 줄 때 얻는 $\frac{2}{3}a$보다 꽤 위에 놓인다.

흔한 실수

- x를 a까지 움직이게 하는 것. $x=a$인 점에서는 $y\le0$이 되어야 한다.
- 홀수 y에 대해 $N=\frac{1}{2}(2a-y)$를 쓰는 것. 이 값은 정수가 아니다.
- 두 경우로 나뉜 주변분포로부터 $\mathrm{E}(Y)$를 구하려는 것. 이 방법도 되기는 하지만, 결합분포의 열을 따라 더하면 홀짝에 따른 경우 나누기를 아예 피할 수 있다.

STEP 너머

결합분포로부터 주변분포를 얻는다는 것은 다른 변수에 대해 합을 취해 없앤다는 뜻이며, 어려운 부분은 언제나 영역의 범위를 정하는 것이다. 연속확률변수에서는 같은 작업이 이

중적분(double integral)의 적분 범위를 정하는 일이 되고, (E(Y)를 구할 때 했던 것처럼) 합의 순서를 바꾸는 것은 적분 순서를 바꾸는 것이 된다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 이산변수와 연속변수 모두에 대해 결합분포와 주변분포를 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.