풀이

STEP II 1988, 문제 15

확률과 통계 • 이산균등분포의 합

> **STEP II 1988, 15번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/88-s2-q15

핵심 아이디어

각 시험지의 점수는 $\{0, 1, \ldots, m\}$ 위에서 균등분포(uniform distribution)를 따르므로, 가능한 모든 점수 목록은 같은 확률로 나타나고, 확률은 조건을 만족하는 목록의 개수를 전체 목록의 개수로 나눈 값일 뿐이다. 시험지가 두 개일 때의 개수는 간단히 나열해서 얻고, 시험지가 세 개일 때의 개수는 세 번째 점수에 대해 조건을 걸고(conditioning) 시험지 두 개일 때의 개수를 더해서 얻는다(이를 **합성곱**(*convolution*)이라 한다). 그러면 대칭성(symmetry, 각 시험지에서 $k \mapsto m - k$)에 의해 가장 높은 범위의 결과는 거의 공짜로 얻어진다.

시작하기 전에 한 가지 주의할 점이 있다. 시험지에 인쇄된 세 시험지에 대한 공식에는 오식(misprint)이 있다. 아래에서 올바른 공식을 유도하고, 인쇄된 공식과 어디가 다른지 짚어 본다.

풀이

X_1, X_2, X_3을 각 시험지의 점수라 하자. 시험지들은 독립적으로 채점되므로 이들은 서로 독립(independent)이고, 각각 $\{0, 1, \ldots, m\}$ 위에서 균등분포를 따른다. 따라서 각 순서쌍 (X_1, X_2)의 확률은 $(m+1)^{-2}$이고, 각 순서 있는 세 쌍의 확률은 $(m+1)^{-3}$이다. $\{0, \ldots, m\}$의 점수로 이루어진 순서쌍(순서 있는 세 쌍) 가운데 합이 n인 것의 개수를 각각 $N_2(n)$, $N_3(n)$으로 나타내자.

시험지가 두 개일 때

$X_1 = k$이고 $X_2 = n - k$이며 둘 다 $[0, m]$에 속해야 하므로

$$\max(0, n - m) \le k \le \min(m, n).$$

- $0 \le n \le m$이면 k는 0부터 n까지이므로 $N_2(n) = n + 1$이다.
- $m < n \le 2m$이면 k는 $n - m$부터 m까지이므로 $N_2(n) = m - (n - m) + 1 = 2m - n + 1$이다.

따라서

$$P(X_1 + X_2 = n) = \begin{cases} \dfrac{n+1}{(m+1)^2}, & 0 \le n \le m, \\ \dfrac{2m-n+1}{(m+1)^2}, & m < n \le 2m, \end{cases}$$

이고, 이것이 $m < n \le 2m$일 때 요구된 결과이며 $0 \le n \le m$일 때의 결과도 함께 얻었다. 이 분포는 "천막" 모양으로, $n = m$까지 한 단계마다 1씩 올라갔다가 그다음부터 내려간다.

시험지가 세 개일 때

세 번째 점수에 대해 조건을 걸자. $X_3 = k$이면 $X_1 + X_2 = n - k$이어야 하므로

$$N_3(n) = \sum_{k=0}^{m} N_2(n-k) = \sum_{j=n-m}^{n} N_2(j),$$

이다. 여기서 $j = n - k$이고, $j < 0$ 또는 $j > 2m$일 때 $N_2(j) = 0$이다.

$0 \le n \le m$인 경우. $j = 0, 1, \ldots, n$만 기여하고(음수인 j는 0을 준다), 이들은 모두 천막의 올라가는 부분에 놓인다.

$$N_3(n) = \sum_{j=0}^{n} (j+1) = \frac{(n+1)(n+2)}{2}, \qquad P(X_1 + X_2 + X_3 = n) = \frac{(n+1)(n+2)}{2(m+1)^3}.$$

$m < n \le 2m$인 경우. 이제 $1 \le n - m \le m < n \le 2m$이므로 범위 $j = n - m, \ldots, n$은 천막의 꼭대기를 가로지른다. 그 지점에서 합을 나눈다.

$$N_3(n) = \sum_{j=n-m}^{m} (j+1) + \sum_{j=m+1}^{n} (2m - j + 1).$$

첫 번째 합은 $n - m + 1$부터 $m + 1$까지의 정수를 더한 것이다.

$$\sum_{j=n-m}^{m} (j+1) = \frac{(m+1)(m+2) - (n-m)(n-m+1)}{2}.$$

두 번째 합에서 $i = 2m - j + 1$로 놓으면, 이 값은 $2m - n + 1$부터 m까지 움직인다.

$$\sum_{j=m+1}^{n} (2m - j + 1) = \frac{m(m+1) - (2m-n)(2m-n+1)}{2}.$$

네 개의 곱을 전개하면

$$\begin{aligned} 2N_3(n) &= (m^2 + 3m + 2) - (n^2 - 2mn + m^2 + n - m) \\ &\quad + (m^2 + m) - (4m^2 - 4mn + n^2 + 2m - n) \\ &= 6mn - 3m^2 - 2n^2 + 3m + 2. \end{aligned}$$

같은 확률로 나타나는 $(m+1)^3$개의 세 쌍으로 나누면

$$P(X_1 + X_2 + X_3 = n) = \frac{6mn - 3m^2 - 2n^2 + 3m + 2}{2(m+1)^3}, \qquad m < n \le 2m.$$

? 생각해 보기 이것은 문제에 인쇄된 공식이 아니다. 누가 옳은가?

인쇄된 공식에는 $-3m^2$ 대신 $-4m^2$이, $(m+1)^3$ 대신 $(m+1)^2$이 있는데, 둘 다 오식이다. 한 가지 경우만 확인하면 결론이 난다. $m = 1, n = 2$일 때 $\{0, 1\}$에서 뽑은 세 점수의 합이 2가 되는 경우는 같은 확률로 나타나는 8가지 중 정확히 3가지이므로 확률은 $\frac{3}{8}$이다. 우리 공식은 $\frac{12-3-8+3+2}{16} = \frac{3}{8}$을 주지만, 인쇄된 공식은 $\frac{12-4-8+3+2}{8} = \frac{5}{8}$을 준다. 어쨌든 분모는 $(m+1)^3$이어야 하는데, 이것이 같은 확률로 나타나는 결과의 개수이기 때문이다. 추가로 확인해 보면, $n = m$일 때 우리의 가운데 공식은 $\frac{(m+1)(m+2)}{2(m+1)^3}$을 주어 첫 번째 경우와 일치한다. 이때 두 합이 이어지므로 당연히 그래야 한다.

$2m < n \le 3m$인 경우. 각 점수 X_i를 $m - X_i$로 바꾸자. 이것 역시 $\{0, \ldots, m\}$ 위에서 균등분포를 따르고, 새로운 세 점수도 여전히 서로 독립이므로

$$\begin{aligned} P(X_1 + X_2 + X_3 = n) &= P\big((m - X_1) + (m - X_2) + (m - X_3) = n\big) \\ &= P(X_1 + X_2 + X_3 = 3m - n). \end{aligned}$$

$2m < n \le 3m$이면 $0 \le 3m - n < m$이므로, n을 $3m - n$으로 바꾸어 첫 번째 경우를 적용할 수 있다.

$$P(X_1 + X_2 + X_3 = n) = \frac{(3m - n + 1)(3m - n + 2)}{2(m+1)^3}, \qquad 2m < n \le 3m.$$

(같은 대칭성을 $3m - n$ 대신 $2m - n$으로 적용하면, 시험지 두 개에 대한 $n \le m$일 때의 결과가 $n > m$일 때의 결과로 바뀐다.)

검산

- 작은 경우: $m = 1$일 때 세 공식은 $n = 0, 1, 2, 3$에 대해 $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$을 주는데, 이들의 합은 1이고 공정한 동전을 세 번 던질 때의 이항분포(binomial distribution)와 일치한다.
- 가운데 공식은 $n \mapsto 3m - n$에 대해 대칭이다. $6mn - 3m^2 - 2n^2$에서 n을 $3m - n$으로 바꾸어도 값이 변하지 않으며, 대칭성 논증에 따르면 반드시 그래야 한다.
- 1부터 8까지의 모든 m에 대해 모든 경우를 나열해 보면, 다섯 공식 모두 개수와 정확히 일치한다. 예를 들어 $m = 10$, $n = 15$일 때 두 방법 모두 $\frac{91}{1331}$을 준다.

흔한 실수

- $N_2(j)$가 두 가지 다른 공식을 가진다는 것을 잊고, 가운데 범위에서 합 전체에 걸쳐 $j + 1$을 사용하는 것.
- 인쇄된 가운데 공식에 도달하려고 계산을 "억지로" 맞추는 것. 신중하게 계산한 결과가 주어진 답과 다르다면, 자신을 의심하기 전에 아주 작은 경우를 시험해 보라.

STEP 너머

독립인 확률변수들의 합의 분포는 각 분포의 **합성곱**이며, 우리의 합 $\sum_k N_2(n-k)$가 바로 그것이다. 확률생성함수(probability generating function)를 쓰면 이 과정이 기계적으로 처리된다. 개수 $N_3(n)$은 $(1 + t + \cdots + t^m)^3$에서 t^n의 계수이다. 시험지의 수가 점점 많아지면 분포의 모양은 천막에서 매끄러운 언덕으로 바뀌고 정규곡선(normal curve)에 가까워진다(중심극한정리(central limit theorem)). Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 생성함수와 중심극한정리를 알맞은 수준에서 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.