풀이

STEP I 1994, 문제 13

확률과 통계 • 지시변수

> **STEP I 1994, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/94-s1-q13

핵심 아이디어

개수를 지시변수(indicator variable, 각각 1 또는 0의 값을 갖는 확률변수)들의 합으로 나타내면 어려운 기댓값(expectation) 계산이 쉬워진다. 각 조각들이 독립(independent)이든 아니든 기댓값은 선형(linear)이기 때문이다. 이 문제에서 각 시행은 분명히 독립이 아니지만, 대칭성(symmetry)에 의해 모든 X_i의 기댓값이 같다는 것을 알 수 있다. 그런 다음 X의 분포(distribution)로부터 $\mathrm{E}(X)$를 직접 한 번 더 계산하고, 두 답을 같다고 놓으면 이항계수(binomial coefficient)에 관한 항등식(identity)을 거의 공짜로 얻는다.

풀이

이 풀이 전체에서 $1 \leq n \leq M$이고, $b > a$ 또는 $b < 0$일 때 $\binom{a}{b} = 0$이라는 규약을 사용한다.

소문항 (i)

각 X_i는 i번째로 꺼낸 토큰이 빨간색이면 1, 그렇지 않으면 0이다. 따라서 합 $X_1 + X_2 + \cdots + X_n$에서 꺼낸 빨간 토큰은 각각 정확히 1씩 기여하고, 빨간색이 아닌 토큰은 0을 기여한다. 그러므로 이 합은 꺼낸 n개의 토큰 중 빨간 토큰의 개수, 즉 X를 센다.

소문항 (ii)

X_i는 0과 1의 값만 가지므로

$$\mathrm{E}(X_i) = 0 \cdot \mathrm{P}(X_i = 0) + 1 \cdot \mathrm{P}(X_i = 1) = \mathrm{P}(i\text{번째로 꺼낸 토큰이 빨간색}).$$

생각해 보기 i번째 토큰이 빨간색일 확률은 그 앞에서 무엇을 꺼냈는지에 따라 달라지지 않을까?

조건부(conditional)로 보면 그렇지만, 조건 없이 보면 그렇지 않다. 토큰에 번호를 붙이고 하나씩 모두 꺼낸다고 상상해 보자. M개 토큰의 모든 배열 순서는 똑같이 일어날 가능성이 있다. 따라서 i번째 자리의 토큰은, 1번째 자리의 토큰과 마찬가지로, M개의 토큰 중 어느 것이든 될 가능성이 똑같다. n번 꺼낸 뒤 멈춘다고 해서 $i \le n$인 자리에서 일어나는 일이 달라지지는 않는다. M개의 토큰 중 m개가 빨간색이므로

$$\mathrm{E}(X_i) = \mathrm{P}(X_i = 1) = \frac{m}{M} \qquad \text{각} i = 1, \ldots, n.$$

($i = 2$인 경우에는 직접 계산으로 확인할 수 있다: $\frac{m}{M} \cdot \frac{m-1}{M-1} + \frac{M-m}{M} \cdot \frac{m}{M-1} = \frac{m(M-1)}{M(M-1)} = \frac{m}{M}$.)

소문항 (iii)

소문항 (i)과 기댓값의 선형성(linearity of expectation, 독립성이 전혀 필요 없다)에 의해

$$\mathrm{E}(X) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n) = n \cdot \frac{m}{M} = \frac{mn}{M},$$

이므로 요구된 결과가 성립한다.

소문항 (iv)

꺼낸 n개 토큰의 집합은 크기가 n인 $\binom{M}{n}$개의 부분집합 중 어느 것이든 될 가능성이 똑같다. 빨간 토큰이 정확히 k개이려면 m개의 빨간 토큰 중 k개를, 나머지 $M-m$개 중 $n-k$개를 골라야 하므로

$$\mathrm{P}(X = k) = \frac{\binom{m}{k}\binom{M-m}{n-k}}{\binom{M}{n}}, \qquad k = 0, 1, \ldots, n.$$

이항계수에 대한 규약 덕분에 $k > m$일 때(빨간 토큰이 모자람)나 $n - k > M - m$일 때(나머지 토큰이 모자람) 이 값은 올바르게 0이 된다. 이것이 **초기하분포**(*hypergeometric distribution*)이다.

소문항 (v)

소문항 (iv)로부터 $\mathrm{E}(X)$를 직접 계산한다. $k = 0$인 항은 0이므로

$$\mathrm{E}(X) = \sum_{k=1}^{n} k\,\mathrm{P}(X = k) = \frac{1}{\binom{M}{n}} \sum_{k=1}^{n} k \binom{m}{k}\binom{M-m}{n-k}.$$

이것을 소문항 (iii)의 결과와 같다고 놓으면

$$\sum_{k=1}^{n} k \binom{m}{k}\binom{M-m}{n-k} = \frac{mn}{M}\binom{M}{n}.$$

마지막으로

$$\frac{n}{M}\binom{M}{n} = \frac{n}{M}\cdot\frac{M!}{n!\,(M-n)!} = \frac{(M-1)!}{(n-1)!\,(M-n)!} = \binom{M-1}{n-1},$$

이므로 우변은 $m\binom{M-1}{n-1}$이고, 이것이 바로 요구된 항등식이다.

검산

- $M=5, m=2, n=3$으로 놓자. 소문항 (iv)에서 $\mathrm{P}(X=0,1,2)=\frac{1}{10},\frac{6}{10},\frac{3}{10}$을 얻고, 이들의 합은 1이며, $\mathrm{E}(X)=\frac{6+6}{10}=\frac{6}{5}=\frac{mn}{M}$이다. 항등식은 $1\cdot2\cdot3+2\cdot1\cdot3=12=2\binom{4}{2}$이 된다.
- $n=M$이면(모두 꺼내면) 확실히 $X=m$이고, 실제로 $\mathrm{E}(X)=mM/M=m$이다.
- $M\le 6$인 모든 주머니에 대해 토큰의 모든 배열 순서를 나열하여 $\mathrm{E}(X_i)$, $\mathrm{E}(X)$, $\mathrm{P}(X=k)$를 확인했고, 모든 $M<30$에 대해 (v)의 항등식을 확인했다.

흔한 실수

- $\mathrm{E}(X)=\sum\mathrm{E}(X_i)$가 성립하려면 독립성이 필요하다고 주장하는 것. 그렇지 않다. 기댓값의 선형성은 항상 성립한다.
- 각 시행을 독립으로 취급하여 (iv)에서 이항분포(binomial distribution) $\mathrm{B}(n,\frac{m}{M})$을 쓰는 것. 이렇게 하면 평균은 맞지만 확률은 틀린다. 토큰을 비복원추출(without replacement)로 꺼내기 때문이다.

STEP 너머

지시변수 방법은 확률론에서 가장 유용한 도구 중 하나이다. 이 방법으로 무작위 순열(random permutation)의 고정점(fixed point)의 개수, 동전 던지기에서 나오는 연(run)의 개수, 무작위 그래프(random graph)의 고립된 꼭짓점(isolated vertex)의 개수의 기댓값을 구할 수 있고, $\mathrm{Var}(\sum X_i)=\sum\mathrm{Var}(X_i)+\sum_{i\neq j}\mathrm{Cov}(X_i,X_j)$를 통해 분산(variance)으로도 확장된다. 하나의 양을 두 가지 방법으로 계산하여 조합론적 항등식(combinatorial identity)을 증명하는 방법 역시 그만큼 일반적이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 책 전체에서 지시변수를 사용하며, 두 아이디어가 실제로 활용되는 모습을 보기에 훌륭한 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.