풀이

STEP III 1990, 문제 15

확률과 통계 • 기하분포

> **STEP III 1990, 15번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/90-s3-q15

핵심 아이디어

각 시행에서 A는 확률 $\frac{1}{4}$, B는 확률 $\frac{3}{4}$로 서로 독립(independent)으로 나온다. 따라서 사건 $\{M = m,\ N = n\}$은 하나의 특정한 시행 결과의 나열이고, 그 확률은 곱으로 주어진다. 이 곱은 m에만 의존하는 인수와 n에만 의존하는 인수의 곱으로 나뉜다. 사실상 이것이 거의 모든 일을 해 준다. 주변분포(marginal distribution)는 기하분포(geometric distribution)이고, 두 확률변수는 독립이며, 이후의 모든 확률은 등비급수(geometric series)가 된다.

풀이

한 번의 시행에 대하여 $p = \mathrm{P}(A) = \frac{3}{12} = \frac{1}{4}$, $q = \mathrm{P}(B) = \frac{9}{12} = \frac{3}{4}$로 놓는다. 각 시행은 서로 독립이다.

사건 $\{M = m,\ N = n\}$은 처음 $m + n + 1$번의 시행 결과가

$$\underbrace{A\,A\cdots A}_{m}\ \underbrace{B\,B\cdots B}_{n}\ A,$$

일 때, 그리고 그때에만 일어난다. 따라서 독립성에 의해 $m \geq 0$, $n \geq 1$에 대하여

$$\mathrm{P}(M = m,\ N = n) = \left(\tfrac{1}{4}\right)^m \left(\tfrac{3}{4}\right)^n \tfrac{1}{4} = \left(\tfrac{1}{4}\right)^{m+1} \left(\tfrac{3}{4}\right)^n.$$

이다. (확률 1로 시행의 나열에는 결국 B가 나오고 그 뒤에 A가 나오므로 이 확률들의 합은 1이다. 아래의 주변분포가 이를 확인해 준다.)

소문항 (i)

M의 주변분포를 구하기 위해 $\sum_{n\geq 1}\left(\frac{3}{4}\right)^n = \dfrac{3/4}{1-3/4} = 3$을 이용하여 $n \geq 1$에 대해 더하면

$$\mathrm{P}(M=m) = \left(\tfrac{1}{4}\right)^{m+1}\cdot 3 = \tfrac{3}{4}\left(\tfrac{1}{4}\right)^m, \qquad m = 0,1,2,\ldots$$

이다. N의 주변분포를 구하기 위해 $\sum_{m\geq 0}\left(\frac{1}{4}\right)^{m+1} = \dfrac{1/4}{1-1/4} = \frac{1}{3}$을 이용하여 $m \geq 0$에 대해 더하면

$$\mathrm{P}(N=n) = \tfrac{1}{3}\left(\tfrac{3}{4}\right)^n = \tfrac{1}{4}\left(\tfrac{3}{4}\right)^{n-1}, \qquad n = 1,2,3,\ldots$$

이다. 두 결과 모두 직접 생각해 보아도 타당하다. M은 첫 번째 B가 나오기 전까지 나온 A의 개수이고, N은 B의 연속이 시작된 뒤 첫 번째 A가 나올 때까지의 대기 시간이다. 각각의 합은 1이다.

소문항 (ii)

$|x| < 1$일 때 $\sum_{k\geq 1} kx^{k-1} = \dfrac{1}{(1-x)^2}$를 이용한다(등비급수를 미분하면 된다). 그러면

$$\mathrm{E}(M) = \tfrac{3}{4}\sum_{m\geq 1} m\left(\tfrac{1}{4}\right)^m = \tfrac{3}{4}\cdot\tfrac{1}{4}\cdot\frac{1}{(3/4)^2} = \tfrac{1}{3},$$

$$\mathrm{E}(N) = \tfrac{1}{4}\sum_{n\geq 1} n\left(\tfrac{3}{4}\right)^{n-1} = \tfrac{1}{4}\cdot\frac{1}{(1/4)^2} = 4.$$

이다.

독립성

주변분포를 곱하면 모든 $m \geq 0, n \geq 1$에 대하여

$$\mathrm{P}(M=m)\,\mathrm{P}(N=n) = \tfrac{3}{4}\left(\tfrac{1}{4}\right)^m\cdot\tfrac{1}{4}\left(\tfrac{3}{4}\right)^{n-1} = \left(\tfrac{1}{4}\right)^{m+1}\left(\tfrac{3}{4}\right)^n = \mathrm{P}(M=m,\ N=n)$$

이다. 따라서 M과 N은 **실제로** 독립이다.

? **생각해 보기** A의 연속 길이가 당연히 B의 연속에 영향을 주지 않을까? 왜 그렇지 않은가?

A의 연속은 첫 번째 B에서 끝나고, 그 시행부터 과정은 새로 시작된다. 남은 시행들은 그 이전의 모든 것과 독립이다. B의 연속 길이는 이 이후의 시행들에만 의존하므로, M을 알아도 N에 대해서는 아무것도 알 수 없다. 결합확률(joint probability)의 인수분해(factorisation)는 이 사실을 대수적으로 나타낸 것이다.

N과 M에 관한 확률

정수 $k \geq 1$에 대하여 기하분포의 꼬리(tail)를 더하면

$$\mathrm{P}(N > k) = \sum_{n \geq k+1} \tfrac{1}{4}\left(\tfrac{3}{4}\right)^{n-1} = \tfrac{1}{4} \cdot \frac{(3/4)^k}{1 - 3/4} = \left(\tfrac{3}{4}\right)^k.$$

이다. 이 식은 $\mathrm{P}(N > 0) = 1$도 주는데, $N \geq 1$이므로 이것도 옳다.

$\mathrm{P}(N > M)$을 구하기 위해 M의 값에 대한 조건을 붙이고 독립성을 이용하면

$$\begin{aligned}\mathrm{P}(N > M) &= \sum_{m \geq 0} \mathrm{P}(M = m)\,\mathrm{P}(N > m) = \sum_{m \geq 0} \tfrac{3}{4}\left(\tfrac{1}{4}\right)^m \left(\tfrac{3}{4}\right)^m \\ &= \tfrac{3}{4} \sum_{m \geq 0} \left(\tfrac{3}{16}\right)^m = \tfrac{3}{4} \cdot \tfrac{16}{13} = \tfrac{12}{13}.\end{aligned}$$

이다. $\mathrm{P}(N = M)$의 경우, $N \geq 1$이므로 공통값은 적어도 1이어야 한다.

$$\begin{aligned}\mathrm{P}(N = M) &= \sum_{m \geq 1} \tfrac{3}{4}\left(\tfrac{1}{4}\right)^m \cdot \tfrac{1}{4}\left(\tfrac{3}{4}\right)^{m-1} = \tfrac{3}{64} \sum_{m \geq 1} \left(\tfrac{3}{16}\right)^{m-1} \\ &= \tfrac{3}{64} \cdot \tfrac{16}{13} = \tfrac{3}{52}.\end{aligned}$$

마지막으로 세 사건 $N > M$, $N = M$, $N < M$은 서로 배반(exclusive)이고 합치면 전체를 이루므로(exhaustive)

$$\mathrm{P}(N < M) = 1 - \tfrac{12}{13} - \tfrac{3}{52} = \tfrac{52-48-3}{52} = \tfrac{1}{52},$$

이고, 이것이 구하려던 결과이다.

검산

- 평균(mean)은 직관과 맞는다. A는 드물게 나오므로 처음 A의 연속은 짧고(평균 $\frac{1}{3}$), B의 연속을 끝낼 A가 나오기까지는 평균 $1/p = 4$번의 시행을 기다린다.
- 직접 계산하면 $\mathrm{P}(M > n) = \left(\tfrac{1}{4}\right)^{n+1}$이므로 $\mathrm{P}(N < M) = \sum_{n \geq 1} \tfrac{1}{4}\left(\tfrac{3}{4}\right)^{n-1}\left(\tfrac{1}{4}\right)^{n+1} = \tfrac{1}{64}\sum_{n \geq 1}\left(\tfrac{3}{16}\right)^{n-1} = \tfrac{1}{64} \cdot \tfrac{16}{13} = \tfrac{1}{52}$이고, 이는 빼서 구한 값과 일치한다. $\mathrm{P}(N > k) = \left(\tfrac{3}{4}\right)^k$인 직접적인 이유는, 첫 번째 B 다음의 k번의 시행이 모두 B여야 하기 때문이다.
- 결합분포를 정확히 더하면 $\mathrm{P}(N > M) = 0.92308$, $\mathrm{P}(N = M) = 0.05769$, $\mathrm{P}(N < M) = 0.01923$이다. 200 000회의 시행 나열에 대한 모의실험(simulation)에서는 0.923, 0.058, 0.019를 얻었고, 표본평균은 0.334와 3.99였다.

흔한 실수

- $\mathrm{P}(N = M)$을 구하는 합에 $m = 0$을 포함하는 것. N은 결코 0이 될 수 없다.

- M이 0일 수 있다는 것(첫 번째 시행이 B일 수 있다)을 잊는 것. 그러면 M의 주변분포가 밀려서 틀린 평균을 얻는다.

STEP 너머

여기서 연속 길이들이 독립인 이유는, 독립시행의 나열이 첫 번째 B처럼 과거에 의해서만 결정되는 임의의 시각에서 "다시 시작"하기 때문이다. 이것을 **강한 마르코프 성질**(strong Markov property)이라고 하며, 이는 재생 이론(renewal theory)과 마르코프 연쇄(Markov chain) 분석의 바탕이 된다. 이들은 대학 1, 2학년에서 배우는 주제이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 기하분포, 독립성, 마르코프 연쇄를 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.