풀이

STEP III 견본 문제지, 문제 16

확률과 통계 • 음이항분포

> **STEP III 견본 문제지, 16번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/spec-s3-q16

핵심 아이디어

아들 r명을 기다리는 것은 아들 한 명을 기다린 다음, 다시 처음부터 시작하여 다음 아들을 기다리는 일을 r번 되풀이하는 것과 같다. 따라서 C는 똑같은 대기 시간(waiting time) r개의 합으로 나뉘고, 나머지는 기댓값의 선형성(linearity of expectation)이 해결해 준다. 소문항 (ii)의 요령은 대통령이 멈춘 뒤에도 출생이 계속 이어진다고 상상하는 것이다. 그러면 주어진 시점까지 대통령이 멈추었는지는 **고정된** 수의 아이들에 관한 질문이 되고, 아들과 딸 사이의 대칭성(symmetry)이 이를 해결한다.

풀이

이 풀이 전체에서 각 출생은 다른 모든 출생과 독립적으로(independently) 확률 $\frac{1}{2}$로 아들이다.

소문항 (i)

$i = 1, 2, \ldots, r$에 대하여, X_i를 $(i-1)$번째 아들 다음부터 i번째 아들까지(그 아들 포함) 태어난 아이의 수라 하자. ($i = 1$이면 처음부터 센다.) 그러면

$$C = X_1 + X_2 + \cdots + X_r.$$

각 X_i는 새로 시작하는 독립적이고 공정한 출생의 연속에서 첫 아들까지의 출생 수를 세므로, 모든 X_i는 X와 같은 분포(distribution)를 가진다. 여기서

$$\mathrm{P}(X = k) = \left(\tfrac{1}{2}\right)^{k-1} \cdot \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^{k}, \qquad k = 1, 2, 3, \ldots,$$

인데, $X = k$는 딸 $k-1$명 다음에 아들이 태어난다는 뜻이기 때문이다.

$\mathrm{E}(X)$를 구하기 위해 $|x| < 1$에서 기하급수(geometric series) $\sum_{k\geq 0} x^k = \dfrac{1}{1-x}$를 미분하여 $\sum_{k\geq 1} kx^{k-1} = \dfrac{1}{(1-x)^2}$을 얻는다. 양변에 x를 곱하고 $x = \frac{1}{2}$를 대입하면

$$\mathrm{E}(X) = \sum_{k=1}^{\infty} k\left(\tfrac{1}{2}\right)^k = \frac{\frac{1}{2}}{\left(1-\frac{1}{2}\right)^2} = 2.$$

이는 납득할 만하다. 모든 아기의 절반이 아들이라면, 아들 한 명을 얻기까지 출생 두 번을 기다릴 것으로 기대된다.

기댓값의 선형성에 의하여 (X_i들이 독립이든 아니든 성립하지만, 여기서는 실제로 독립이다)

$$\mathrm{E}(C) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_r) = 2r.$$

소문항 (ii)

끝없이 이어지는 아기들의 성별이 미리 독립적이고 공정하게 정해져 있고, 대통령은 이 수열에서 r번째 아들에 이를 때까지 아기를 낳는다고 상상하자. 이렇게 해도 C에 관해서는 아무것도 달라지지 않지만, 대통령이 더 일찍 멈추는 경우에도 "처음 $2r-1$명의 아기"에 대하여 이야기할 수 있게 된다.

> **생각해 보기** 사건 $C < 2r$를 수열의 처음 $2r-1$명의 아기만을 써서 어떻게 나타낼 수 있을까?

C는 정수이므로 $C < 2r$는 $C \leq 2r-1$, 즉 r번째 아들이 $2r-1$번째 자리 또는 그 이전에 태어난다는 뜻이다. 이는 처음 $2r-1$명의 아기 중 아들이 적어도 r명일 때, 그리고 그때에만 일어난다. 따라서

$$\mathrm{P}(C < 2r) = \mathrm{P}(\text{아들 수가 적어도}r\text{명 (처음}2r-1\text{명의 아기 중)}).$$

이제 B와 G를 이 $2r-1$명의 아기 중 아들과 딸의 수라 하면 $B+G = 2r-1$이다. $2r-1$은 홀수이므로 B와 G는 같을 수 없고, 둘 중 정확히 하나만 r 이상이다.

- $B \geq r$이면 $G = 2r-1-B \leq r-1$이다.
- $B \leq r-1$이면 $G \geq r$이다.

따라서 사건 $\{B \geq r\}$와 $\{G \geq r\}$는 서로 배반(mutually exclusive)이고 둘을 합하면 모든 경우를 포함하므로

$$\mathrm{P}(B \geq r) + \mathrm{P}(G \geq r) = 1.$$

그런데 매 출생에서 아들과 딸은 같은 확률로 태어나므로 "아들"과 "딸"이라는 이름표를 서로 바꾸어도 어떤 확률도 달라지지 않으며, $\mathrm{P}(B \geq r) = \mathrm{P}(G \geq r)$이다. 그러므로 각각은

$\frac{1}{2}$이고

$$\mathrm{P}(C < 2r) = \tfrac{1}{2},$$

이 되어 증명이 끝난다.

따라서 가족 수의 기댓값은 $2r$이지만, 가족 수가 $2r$보다 작은 경우는 정확히 절반이다. 가족 수의 분포는 한쪽으로 치우쳐(skewed) 있다. $2r$보다 조금 작은 값이 흔하고, 가끔 나타나는 매우 긴 딸의 연속이 평균을 끌어올린다.

검산

- $r = 1$일 때: $\mathrm{E}(C) = 2$이고, $C < 2$는 첫 아기가 아들이라는 뜻이므로 그 확률은 $\frac{1}{2}$이다.
- $r = 2$일 때: $C < 4$는 $C = 2$ 또는 3이라는 뜻이고, 그 확률은 각각 $\frac{1}{4}$와 $2 \cdot \frac{1}{8} = \frac{1}{4}$이므로 합은 $\frac{1}{2}$이다.
- $r = 1, \ldots, 8$에 대하여 $c\binom{c-1}{r-1}2^{-c}$를 정확히 합하면 매번 $\mathrm{E}(C) = 2r$이고 $\mathrm{P}(C < 2r) = \frac{1}{2}$이다. $r = 3$이고 가족 200 000개로 한 시뮬레이션(simulation)에서는 평균이 6.00, $\mathrm{P}(C < 6) \approx 0.499$였다.

흔한 실수

- "아이들의 절반이 아들이므로 $C = 2r$이다"라고 주장하는 것. 평균적으로는 맞지만 C 자체는 확률적으로 정해진다. 소문항 (i)에는 단순한 추측이 아니라 대기 시간으로의 분해가 필요하다.
- 소문항 (ii)에서 대통령의 실제 자녀만 고려하는 것. $C < 2r - 1$일 때는 $2r - 1$명의 자녀가 **존재하지 않는다**. 멈추는 시점 이후로 출생의 수열을 연장하는 것이 바로 세기 논법을 타당하게 만든다.
- $2r - 1$이 홀수라는 사실을 잊는 것. 바로 이 점이 아들과 딸의 수가 같아지는 경우를 배제하고 두 사건을 서로의 여사건(complementary event)으로 만든다.

STEP 너머

C의 분포는 **음이항분포**(negative binomial distribution)이다. 즉 $c \geq r$에 대하여 $\mathrm{P}(C = c) = \binom{c-1}{r-1}2^{-c}$이다. 확률변수(random variable)를 더 간단한 조각들의 합으로 쓰고 기댓값의 선형성을 이용하는 것은 대학 1학년 확률론에서 가장 강력한 도구 가운데 하나이다. 그리고 소문항 (ii)에서 "대기 시간"을 "고정된 횟수의 시행에서의 개수"와 연결하는 요령은 음이항분포와 이항분포(binomial distribution)를 (그리고 연속 시간에서는 감마분

포(gamma distribution)와 푸아송분포(Poisson distribution)를) 이어 주는 것과 같은 쌍대성(duality)이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 이 두 아이디어를 많은 예와 함께 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.