풀이

STEP III 2001, 문제 12

확률과 통계 • 지시변수를 이용한 기댓값

> **STEP III 2001, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/01-s3-q12

핵심 아이디어

소문항 (i)에서는 검은 공의 개수의 분포를 구하려 하지 말고, 그 개수를 검은 공 하나마다 지시확률변수(indicator) 하나씩으로 쪼갠 뒤 기댓값의 선형성(linearity of expectation)을 이용한다. 주머니가 빌 때까지 비복원추출(drawing without replacement)을 하는 것은 모든 공을 무작위로 한 줄로 늘어놓는 것과 같으므로, 각 지시확률변수는 대칭성 논증으로 해결된다. 소문항 (ii)에서는 흰 공이 뽑히기 전까지 주머니의 구성이 전혀 바뀌지 않으므로, 각 흰 공을 기다리는 시간은 기하분포(geometric distribution)를 따른다. 그리고 "이를 이용하여" 부분은 흰 공 하나가 빠지고 나면 흰 공이 하나 적은 같은 상황으로 돌아온다는 점을 알아차리는 데서 나온다.

풀이

소문항 (i)

주머니가 빌 때까지 계속 공을 꺼낸다고 상상하자. 매번 무작위로 꺼내고 아무것도 되돌려 넣지 않으므로, $b+w$개의 공이 나오는 순서는 이 공들의 균등하게 무작위인 순열이다. 마지막 흰 공이 꺼내질 때까지 뽑힌 검은 공은 정확히 이 순서에서 마지막 흰 공보다 앞에 오는 검은 공들이다.

검은 공에 $1, 2, \ldots, b$의 번호를 붙이고, 검은 공 j가 마지막 흰 공보다 먼저 나오면 $I_j = 1$, 그렇지 않으면 $I_j = 0$이라 하자. 마지막 흰 공이 꺼내질 때까지 뽑힌 검은 공의 개수를 N이라 하면,

$$N = I_1 + I_2 + \cdots + I_b, \qquad \text{따라서} \qquad \mathrm{E}(N) = \sum_{j=1}^{b} \mathrm{P}(I_j = 1)$$

이다. 이는 기댓값의 선형성에 의한 것이며, 여기에는 I_j들 사이의 독립성이 전혀 필요하지 않다.

생각해 보기 검은 공 j가 모든 흰 공보다 뒤에 나올 확률은 얼마인가?

검은 공 j와 w개의 흰 공으로 이루어진 $w+1$개의 공만 살펴보자. 모든 공의 균등하게 무작위인 순열에서 이 $w+1$개 공의 상대적인 순서 역시 균등하게 무작위이므로, 검은 공 j는 이들 사이의 $w+1$개 자리 중 어느 자리에나 같은 확률로 놓인다. 검은 공 j가 이 무리에서 맨 마지막일 확률은 $\frac{1}{w+1}$이고, 이것이 바로 사건 $I_j = 0$이다. 따라서

$$\mathrm{P}(I_j = 1) = 1 - \frac{1}{w+1} = \frac{w}{w+1},$$

이고

$$\boxed{\mathrm{E}(N) = \frac{bw}{w+1}.}$$

이다.

소문항 (ii)

이제 검은 공은 다시 주머니에 넣으므로, 흰 공이 뽑힐 때까지 주머니에는 언제나 검은 공 b개와 흰 공 w개가 들어 있다. 주머니의 구성이 매번 같으므로, 각 시행은 이전 시행과 독립적으로 확률 $p = \frac{b}{b+w}$로 검은 공, 확률 $q = \frac{w}{b+w}$로 흰 공이 나온다.

첫 번째 흰 공이 나오기 전까지 검은 공이 뽑힌 횟수를 X라 하자. $X = n$은 검은 공이 n번 나온 뒤 흰 공이 나온다는 뜻이므로

$$\mathrm{P}(X = n) = p^n q, \qquad n = 0, 1, 2, \ldots$$

이다. $b = 0$이면 $X = 0$이므로 결과는 자명하다. 따라서 $b \geq 1$이라 하면 $0 < p < 1$이고, $|x| < 1$일 때 $\sum_{n\geq 1} n x^{n-1} = \frac{1}{(1-x)^2}$(등비급수를 미분하면 얻는다)를 이용하면

$$\mathrm{E}(X) = \sum_{n=0}^{\infty} n p^n q = pq \sum_{n=1}^{\infty} n p^{n-1} = \frac{pq}{(1-p)^2} = \frac{p}{q} = \frac{b/(b+w)}{w/(b+w)} = \frac{b}{w},$$

이다. 여기서 $1 - p = q$를 이용하였다. 이것이 구하려던 결과이다.

마지막 흰 공에 대해서는, 검은 공이 뽑힌 전체 횟수를 $X_w + X_{w-1} + \cdots + X_1$로 쓴다. 여기서 X_k는 주머니에 흰 공이 정확히 k개 남아 있는 동안 검은 공이 뽑힌 횟수이다. 흰 공이 k개 남아 있는 동안 주머니에는 검은 공 b개와 흰 공 k개가 들어 있으므로, 이 단계는 w를 k로 바꾼 첫 번째 흰 공 문제와 정확히 같고, 방금 증명한 결과에 의해 $\mathrm{E}(X_k) = \frac{b}{k}$이다. 기댓값의 선형성에 의해 마지막 흰 공이 꺼내질 때까지 검은 공이 뽑힌 횟수의 기댓값은

$$\boxed{\sum_{k=1}^{w} \frac{b}{k} = b\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{w}\right).}$$

이다.

검산

$w = 1$이면 두 소문항 모두 직접 쉽게 확인할 수 있다. (i)에서 각 검은 공은 하나뿐인 흰 공보다 확률 $\frac{1}{2}$로 앞에 오므로 $\frac{b}{2}$를 얻고, 이는 $\frac{bw}{w+1}$과 일치한다. (ii)에서는 단계가 하나뿐이므로 합은 그냥 $\frac{b}{1} = b$이다. (i)에서 $w \to \infty$이면 거의 모든 검은 공이 마지막 흰 공보다 먼저 나오는데, 실제로 $\frac{bw}{w+1} \to b$이다.

(i)은 모든 $b \le 4$, $w \le 4$에 대해 가능한 순서를 모두 나열하여 확인하였고, $b = 3$, $w = 4$일 때 정확히 $\frac{12}{5}$를 얻는다. (ii)는 $b = 3$, $w = 4$로 2×10^5회 모의실험을 한 결과 평균이 약 6.245였으며, 이는 $3(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}) = \frac{25}{4} = 6.25$와 비교된다.

흔한 실수

- (i)에서 뽑힌 검은 공들이 서로 독립이라고 가정하는 것. 기댓값의 선형성에는 독립성이 필요하지 않으며, 바로 그 때문에 지시확률변수가 이토록 잘 통한다.
- (ii)에서 흰 공이 뽑히는 시행까지 함께 세는 기하분포를 사용하는 것. 그러면 $\frac{1}{q} = \frac{b+w}{w}$를 얻는데, 이는 답보다 1만큼 크다.
- 마지막 부분의 모든 단계에서 분모에 $b + w$를 쓰는 것. 주머니 속 흰 공의 개수는 단계마다 하나씩 줄어든다.

STEP 너머

두 번째 소문항은 모습을 바꾼 쿠폰 수집가 문제(coupon collector's problem)이다. 단계별 기하분포 대기 시간을 모두 더하면 조화급수(harmonic sum)가 나오며, 이는 $\ln w$처럼 증가한다. 소문항 (i)에서처럼 개수를 지시확률변수의 합으로 나타내는 것은 확률론에서 가장 강력한 기법 가운데 하나이며, 무작위 순열(random permutation)의 고정점(fixed point), 기록(record), 순환(cycle)의 개수의 기댓값을 계산할 때 다시 만나게 될 것이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에는 두 아이디어의 예가 많이 실려 있다.

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.