풀이

STEP I 2013, 문제 12

확률과 통계 • 비복원추출

> **STEP I 2013, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/13-s1-q12

핵심 아이디어

알약 k개를 "무작위로" 고른다는 것은 용기에 남아 있는 알약 k개짜리 집합이 모두 같은 가능성으로 뽑힌다는 뜻이다. 따라서 각 날의 확률(probability)은 개수의 비, 곧 조건에 맞는 선택의 수를 $\binom{\text{남은 알약 수}}{k}$로 나눈 값이다. 마지막 날에 관한 질문의 요령은 대칭성(symmetry)이다. 마지막 날을 위해 남는 알약들은 첫째 날에 고른 알약들과 똑같이 무작위이다. (iii)에서는 날들이 깔끔하게 이어지는데, 좋은 날을 보내고 나면 용기의 상태가 더 짧은 휴가의 처음 상태와 정확히 같아지기 때문이다.

풀이

알약들은 겉보기에 똑같지만, 풀이 내내 서로 구별할 수 있다고(예를 들어 번호가 매겨져 있다고) 생각한다. 이렇게 해도 어떤 확률도 바뀌지 않으며, N개에서 k개를 고르는 $\binom{N}{k}$가지 선택이 모두 같은 가능성을 갖게 된다.

소문항 (i)

$k = 3$, $n = 3$이면 알약은 9개이고 종류마다 3개씩 있다. 이 중 3개를 고르는 방법의 수는 $\binom{9}{3} = 84$이다. 종류마다 하나씩 얻으려면 종류 1의 알약 3개 중 하나, 종류 2의 3개 중 하나, 종류 3의 3개 중 하나를 골라야 하므로 그 방법은 $3^3 = 27$가지이다. 따라서

$$\mathrm{P}(\text{첫째 날 종류별로 하나씩}) = \frac{27}{84} = \frac{9}{28},$$

이고, 이것이 구하려던 값이다.

생각해 보기 마지막 날의 확률을 구할 때 왜 첫째 날과 둘째 날에 일어나는 일을 따져 볼 필요가 없을까?

휴가 동안 아홉 개의 알약은 세 개씩 세 묶음, 곧 첫째 날 묶음, 둘째 날 묶음, 셋째 날 묶음(남은 것)으로 무작위로 나뉜다. 묶음을 1, 2, 3의 순서로 고르든 3, 2, 1의 순서로 고르든 분포는 같다. 알약들을 이름 붙은 세 묶음으로 나누는 각각의 방법이 모두 같은 가능성을 갖기 때문이다. 특히 셋째 날을 위해 남는 집합은 첫째 날에 고른 집합과 똑같이, 알약 세 개로 이루어진 균등하게 무작위인(uniformly random) 집합이다. 따라서 마지막 날에 종류별로 하나씩 남을 확률도

$$\frac{9}{28}.$$

이다.

소문항 (ii)

이제 알약은 $3n$개이고 종류마다 n개씩 있다. 같은 가능성의 선택이 $\binom{3n}{3}$가지 있고, 그중 n^3가지가 종류별로 하나씩을 포함한다. 따라서

$$\begin{aligned}\mathrm{P}(\text{첫째 날 종류별로 하나씩}) &= \frac{n^3}{\binom{3n}{3}} = \frac{6n^3}{3n(3n-1)(3n-2)} \\ &= \frac{2n^2}{(3n-1)(3n-2)}.\end{aligned}$$

$n = 3$이면 이 값은 $\frac{18}{8\cdot 7} = \frac{9}{28}$로, (i)과 일치한다.

소문항 (iii)

이제 $k = 2$이고 알약은 $2n$개이며, 두 종류가 각각 n개씩 있다. $m = n, n-1, \ldots, 1$에 대하여, 하루를 시작할 때 용기에 종류마다 m개씩 알약이 있는 날에 내가 종류별로 하나씩 고르는 사건을 G_m이라 하자(따라서 첫째 날에는 $m = n$이다).

용기에 종류마다 m개씩 있으면 알약 두 개를 고르는 방법은 $\binom{2m}{2} = m(2m-1)$가지이고, 그중 $m \cdot m$가지가 종류별로 하나씩이다. 따라서

$$\mathrm{P}(\text{종류별로 하나씩} \mid m\text{개씩 남음}) = \frac{m^2}{m(2m-1)} = \frac{m}{2m-1}.$$

지금까지의 모든 날이 좋은 날이었다면, 매일 종류마다 정확히 하나씩 빠졌으므로 다음 날은 종류마다 $m-1$개씩으로 시작한다. 조건부 확률(conditional probability)을 날마다 곱하면

$$\mathrm{P}(\text{매일 종류별로 하나씩}) = \prod_{m=1}^{n} \frac{m}{2m-1} = \frac{n!}{1 \times 3 \times 5 \times \cdots \times (2n-1)}.$$

홀수들의 곱을 정리하기 위해 분자와 분모에 짝수들의 곱 $2\times 4\times\cdots\times(2n)=2^n\,n!$을 곱하면

$$1\times 3\times\cdots\times(2n-1)=\frac{(2n)!}{2^n\,n!}.$$

따라서

$$\boxed{\mathrm{P}(\text{매일 종류별로 하나씩})=\frac{2^n\,(n!)^2}{(2n)!}}$$

(검산으로, $n=1$이면 1이 나오는데 이는 옳다. 종류마다 알약이 하나씩뿐이면 둘 다 먹을 수밖에 없다.)

이제 $n!$과 $(2n)!$에 스털링(Stirling) 근사(Stirling's approximation)를 사용한다.

$$(n!)^2\approx 2n\pi\left(\frac{n}{\mathrm{e}}\right)^{2n},$$

$$(2n)!\approx\sqrt{4n\pi}\left(\frac{2n}{\mathrm{e}}\right)^{2n}=2\sqrt{n\pi}\,2^{2n}\left(\frac{n}{\mathrm{e}}\right)^{2n}.$$

인수 $(n/\mathrm{e})^{2n}$이 약분되어

$$\frac{2^n\,(n!)^2}{(2n)!}\approx\frac{2^n\cdot 2n\pi}{2\sqrt{n\pi}\,2^{2n}}=2^{-n}\,\frac{n\pi}{\sqrt{n\pi}}=2^{-n}\sqrt{n\pi},$$

가 남고, 이것이 구하려던 근사식이다.

검산

- 알약을 뽑는 모든 방법을 컴퓨터로 빠짐없이 세어 보면, (i)에서 첫째 날과 셋째 날 모두 $\frac{9}{28}$임이 확인되고, $n\le 7$에 대하여 (ii)의 공식이, $n\le 4$에 대하여 (iii)의 공식이 확인된다.
- $n=10$일 때 (iii)의 정확한 확률은 $\frac{2^{10}(10!)^2}{20!}\approx 0.005542$이고, $2^{-10}\sqrt{10\pi}\approx 0.005474$이므로 근삿값이 약 1.3% 작다. 스털링 근사에서 예상되듯이 n이 커지면 이 비는 1로 다가간다($n=50$에서 약 1.0025이다).
- (iii)을 다르게 세는 방법: $2n$개의 알약을 n개의 쌍으로 이루어진 순서 있는 목록으로 나누는 방법은 $(2n)!/2^n$가지이고 모두 같은 가능성을 갖는다. 좋은 분할은 날마다 종류 1의 알약과 종류 2의 알약을 짝짓는 것으로 $n!\times n!$가지이다. 그 비는 역시 $2^n(n!)^2/(2n)!$이다.

흔한 실수

- 알약을 하나씩 뽑는다고 보고 $\frac{3}{9}\times\frac{3}{8}\times\frac{3}{7}$이라고 쓰는 것. 이는 종류 1, 2, 3이 그 순서대로 나올 확률이다. 종류들은 3!가지 순서로 나올 수 있으므로 $6\times\frac{27}{504}=\frac{9}{28}$이 된다. (또는 첫 알약은 아무것이나 될 수 있으므로 $1\times\frac{6}{8}\times\frac{3}{7}=\frac{9}{28}$이다.)

- 마지막 날의 확률을 구하려고 첫째 날에 일어나는 모든 일에 대해 조건을 걸어 따지는 것. 그렇게 해도 답은 나오지만 길고, 대칭성을 쓰면 답이 곧바로 나온다.
- $(2n)!$에 스털링 근사를 쓸 때 제곱근 안을 포함하여 모든 곳의 n을 $2n$으로 바꾸어야 한다. $\sqrt{2(2n)\pi} = 2\sqrt{n\pi}$이지 $\sqrt{2n\pi}$가 아니다.

STEP 너머

(iii)의 답은 $2^n/\binom{2n}{n}$이고, 근사식 $\binom{2n}{n} \approx 4^n/\sqrt{n\pi}$는 확률론 곳곳에 등장한다. 이 근사식은 단순 확률보행(simple random walk)이 $2n$걸음 뒤에 출발점으로 돌아와 있을 확률을 주며, 1차원이나 2차원에서 그러한 보행이 확률 1로 출발점에 돌아온다는 증명(proof)의 원동력이 된다. (i)의 대칭성 논증은 **교환가능성**(exchangeability)이라는 개념이다. 즉 무작위 추출을 하는 순서는 그 결합분포(joint distribution)를 바꾸지 않는다. 두 개념 모두 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문), 그리고 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에서 전개된다.

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.