풀이

STEP I 1998, 문제 13

확률과 통계 • 복원추출

> **STEP I 1998, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/98-s1-q13

핵심 아이디어

이것은 조건부 확률(conditional probability) 문제이다. 조건으로 주어진 사건은 "n번 꺼내는 동안 파스틸(pastille)을 정확히 하나 먹었다"는 것이고, 이 사건을 그 파스틸을 **언제** 먹었는지에 따라 나눈다. 봉지 안의 내용물은 빨간 파스틸을 먹는 순간에만 바뀌므로, 나눈 각 조각은 간단한 확률들의 곱이 된다. 조각들을 더하면 등비급수(geometric series)가 되고, 바로 여기서 두 n제곱의 차가 나온다.

풀이

처음에 들어 있는 파스틸의 개수를 $N = r + b$라 하자. $r \geq 1$이고 $b \geq 1$이라고 가정한다($b = 0$이고 $n \geq 2$이면 파스틸을 정확히 하나만 먹는 것은 불가능하므로 조건으로 삼을 사건이 없다).

매번 꺼낼 때의 봉지. 빨갛지 않은 파스틸은 다시 넣으므로, 봉지는 빨간 파스틸을 먹을 때에만 바뀐다. 빨간 파스틸을 먹기 전에는 봉지에 파스틸이 N개 있고 그중 b개가 빨갛지 않다. 먹은 뒤에는 파스틸이 $N-1$개 있고, 빨갛지 않은 것은 여전히 b개이다. 매번 꺼내기 전에 봉지를 충분히 흔들므로, 각 시행은 그 순간 봉지에 들어 있는 것들 가운데에서 새로 균등하게 하나를 고르는 것이다. 따라서 확률을 한 번 꺼낼 때마다 차례로 곱할 수 있다.

먹은 시점에 따라 나누기. $k = 1, 2, \ldots, n$에 대하여 E_k를 n번의 시행에서 먹은 유일한 파스틸을 k번째 시행에서 먹은 사건이라 하자. 그러면 $1, \ldots, k-1$번째 시행에서는 가득 찬 봉지에서 빨갛지 않은 것이 나오고, k번째 시행에서는 가득 찬 봉지에서 빨간 것이 나오며,

$k+1, \ldots, n$번째 시행에서는 줄어든 봉지에서 빨갛지 않은 것이 나온다.

$$\mathrm{P}(E_k) = \left(\frac{b}{N}\right)^{k-1} \frac{r}{N} \left(\frac{b}{N-1}\right)^{n-k} = \frac{r\,b^{n-1}}{N^k\,(N-1)^{n-k}}.$$

사건 $E_1, \ldots, E_n$은 서로 배반(disjoint)이고, 이들의 합집합은 파스틸을 정확히 하나 먹은 사건 E이다. 따라서

$$\mathrm{P}(E) = r\,b^{n-1} \sum_{k=1}^{n} \frac{1}{N^k (N-1)^{n-k}} = \frac{r\,b^{n-1}}{N^n (N-1)^n} \sum_{k=1}^{n} N^{n-k} (N-1)^k,$$

이다. 여기서 각 항의 분자와 분모에 $N^n(N-1)^n$을 곱하였다.

? **생각해 보기** $\sum_{k=1}^{n} N^{n-k}(N-1)^k$를 어떻게 더할까?

이것은 첫째항이 $N^{n-1}(N-1)$이고 공비가 $\frac{N-1}{N}$인 등비급수이다. 더 빠르게는, 잘 알려진 인수분해(factorisation)

$$x^n - y^n = (x-y)(x^{n-1} + x^{n-2}y + \cdots + y^{n-1})$$

에 $x = N, y = N-1$을 대입하면 $x - y = 1$이므로

$$\sum_{k=1}^{n} N^{n-k}(N-1)^k = (N-1) \sum_{j=0}^{n-1} N^{n-1-j}(N-1)^j = (N-1)(N^n - (N-1)^n).$$

그러므로

$$\mathrm{P}(E) = \frac{r\,b^{n-1}(N^n - (N-1)^n)}{N^n (N-1)^{n-1}}.$$

조건부 확률. "마지막 시행에서 먹었고, 하나만 먹었다"는 사건은 E_n이며, $\mathrm{P}(E_n) = \frac{r\,b^{n-1}}{N^n}$이다. $E_n \subseteq E$이므로

$$\begin{aligned}\mathrm{P}(E_n \mid E) = \frac{\mathrm{P}(E_n)}{\mathrm{P}(E)} &= \frac{r\,b^{n-1}}{N^n} \cdot \frac{N^n (N-1)^{n-1}}{r\,b^{n-1}(N^n - (N-1)^n)} \\ &= \frac{(N-1)^{n-1}}{N^n - (N-1)^n} = \frac{(r+b-1)^{n-1}}{(r+b)^n - (r+b-1)^n},\end{aligned}$$

이고, 이것이 구하려던 결과이다.

r와 b는 그 합을 제외하면 완전히 약분되어 사라진다는 점에 주목하자. 사실 위의 계산은 더 많은 것을 보여 준다. E가 주어졌을 때 파스틸을 k번째 시행에서 먹었을 확률은 $(\frac{N-1}{N})^k$에 비례하므로, 앞선 시행일수록 더 가능성이 크다. 이는 이치에 맞는다. 일찍 먹으면 봉지에서 빨갛지 않은 것의 비율이 더 커지고, 그러면 남은 시행에서 "더 이상 빨간 것이 나오지 않을" 확률이 커지기 때문이다.

검산

- $n = 1$: 공식은 $\dfrac{1}{N-(N-1)} = 1$을 준다. 유일한 시행이 곧 마지막 시행이므로 옳다.
- $n = 2$: 공식은 $\dfrac{N-1}{2N-1}$을 주는데, 이는 $\frac{1}{2}$보다 조금 작으며 앞선 시행이 약간 더 가능성이 크다는 사실과 들어맞는다.
- $r = 3, b = 5, n = 4$일 때 공식은 $343/1695 \approx 0.2024$를 준다. 네 시행이 모두 똑같이 가능하다면 $\frac{1}{4}$일 것이다. 4×10^5회의 시뮬레이션은 0.2028을 주었고, 봉지의 상태를 추적하며 가능한 모든 시행 순서를 정확히 나열해 보니 $1 \le r, b \le 4, 1 \le n \le 8$인 모든 경우에 공식이 정확히 재현되었다.

흔한 실수

- 빨간 것을 먹은 **뒤의** 빨갛지 않은 시행에도 b/N을 쓰는 것. 이제 봉지에는 파스틸이 $N-1$개 있으며, 이것이 바로 이 문제의 핵심이다. 처음부터 끝까지 b/N을 쓰면 모든 E_k의 확률이 같아져서 답이 $1/n$이 되어 버린다.
- 조건을 거는 것을 잊는 것. 파스틸을 정확히 하나 먹었다는 정보가 주어졌으므로, $\mathrm{P}(E_n) = rb^{n-1}/N^n$ 그 자체는 답이 아니다.

STEP 너머

여기서의 봉지는 간단한 마르코프 연쇄(Markov chain)이다. 그 상태는 남은 빨간 파스틸의 개수이고, 상태가 한 단계 내려갈 확률은 현재 상태에만 의존한다. 이러한 연쇄의 최종 상태를 조건으로 두고 그것이 거쳐 온 경로를 묻는 것이 "은닉 마르코프 모형(hidden Markov model)"과 과거 이력에 대한 베이즈 추론(Bayesian inference)의 바탕이 되는 아이디어이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 조건부 확률과 마르코프 연쇄를 정확히 이 수준에서 다루며, Blitzstein과 Hwang의 *Introduction to Probability*(확률 입문)에는 같은 정신의 문제가 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.