풀이

STEP II 2008, 문제 13

확률과 통계 • 비복원추출

> **STEP II 2008, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/08-s2-q13

핵심 아이디어

모든 말을 추적할 필요는 없다. 검은 말만 추적하면 된다. 각 검은 말의 운명은 두 번의 추출로 결정되며, 각 추출에서 특정한 말 하나가 뽑힐 확률(probability)은 단순히 (뽑는 개수)/(주머니에 든 개수)이다. 첫 번째 추출에서 일어나는 일에 따라 경우를 나누면 두 소문항 모두 짧은 합으로 바뀌고, 둘 다 예상보다 훨씬 간단하게 정리된다.

풀이

이 풀이 전체에서 "무작위로 뽑는다"는 것은 주머니 안의 k개의 말로 이루어진 모든 집합이 뽑힐 가능성이 똑같다는 뜻이다. 반복해서 사용할 사실은 다음과 같다. m개가 든 주머니에서 k개의 말을 무작위로 뽑으면, 대칭성(symmetry)에 의해 **특정한** 말 하나가 그 안에 포함될 확률은 k/m이다(m개의 말 각각이 포함될 가능성은 똑같고, 평균적으로 그중 k개가 포함되기 때문이다).

첫 번째 옮기기가 끝나면 주머니 P에는 $n-k$개, 주머니 Q에는 $n+k$개의 말이 들어 있다. 두 번째 옮기기에서는 그 $n+k$개 가운데 k개를 뽑아 P로 옮긴다.

소문항 (i)

P에서 출발하는 검은 말을 B라 하자. 첫 번째 옮기기에서 B가 움직이는지에 따라 경우를 나눈다.

- B가 P에서 뽑히지 않는 경우: 확률은 $\dfrac{n-k}{n}$이다. 이때 두 번째 옮기기는 Q에서만 말을 꺼내므로 B는 끝까지 P에 남는다.
- B가 P에서 뽑히는 경우: 확률은 $\dfrac{k}{n}$이다. 이제 B는 Q에 있는 $n+k$개의 말 중 하나이고,

다시 뽑혀 돌아오는 k개에 포함될 때에만 P로 돌아온다. 그 확률은 $\dfrac{k}{n+k}$이다.

이 두 경우는 서로 배반이고 모든 경우를 포함하므로

$$\mathrm{P}(B\text{의 최종 위치가}P) = \frac{n-k}{n} + \frac{k}{n}\cdot\frac{k}{n+k} = \frac{(n-k)(n+k)+k^2}{n(n+k)}$$
$$= \frac{n^2}{n(n+k)} = \frac{n}{n+k}.$$

이 값은 k에 대하여 순감소하므로, $0 \le k \le n$에서 최댓값을 갖는 것은 오직 $k=0$일 때이며 그 값은 1이다. (놀랄 일은 아니다. 아무것도 옮기지 않으면 검은 말은 당연히 제자리에 있다.)

소문항 (ii)

이제 P에서 출발하는 검은 말을 B_1, Q에서 출발하는 검은 말을 B_2라 하자. 두 말이 같은 주머니에서 끝나는 사건을 S라 하자. 이번에도 첫 번째 옮기기에서 B_1이 움직이는지에 따라 경우를 나눈다.

경우 1: B_1이 P에 남는 경우 (확률 $\frac{n-k}{n}$). 이때 B_1은 무슨 일이 있어도 P에서 끝나므로, S는 B_2가 Q의 $n+k$개 말 가운데 뽑히는 k개에 포함될 때에만 일어난다.

$$\mathrm{P}(S \mid \text{경우 1}) = \frac{k}{n+k}.$$

경우 2: B_1이 Q로 옮겨지는 경우 (확률 $\frac{k}{n}$). 이제 두 검은 말이 모두 Q의 $n+k$개 가운데 있고, S는 두 번째 추출에서 두 말을 **모두** 뽑거나 **둘 다** 뽑지 않을 때 일어난다.

> **생각해 보기** $n+k$개의 말에서 무작위로 고른 k개짜리 집합이 주어진 두 말을 모두 포함하거나, 둘 다 포함하지 않을 확률은 얼마인가?

부분집합의 개수를 센다. 가능성이 똑같은 선택은 모두 $\binom{n+k}{k}$가지이다. 두 검은 말을 모두 포함하는 선택은 나머지 $n+k-2$개에서 다른 $k-2$개를 고르는 것이고, 둘 다 포함하지 않는 선택은 $n+k-2$개의 흰 말에서 k개를 모두 고르는 것이다. 나누면

$$\mathrm{P}(\text{둘 다}) = \binom{n+k-2}{k-2}\Big/\binom{n+k}{k} = \frac{k(k-1)}{(n+k)(n+k-1)},$$
$$\mathrm{P}(\text{둘 다 아님}) = \binom{n+k-2}{k}\Big/\binom{n+k}{k} = \frac{n(n-1)}{(n+k)(n+k-1)}.$$

(이 식들은 $k=0,1$일 때에도 성립하며, 이때 $\mathrm{P}(\text{둘 다}) = 0$이다.) 따라서

$$\mathrm{P}(S \mid \text{경우 2}) = \frac{k(k-1)+n(n-1)}{(n+k)(n+k-1)}.$$

두 경우 합치기. 전확률 법칙(law of total probability)에 의해

$$\begin{aligned}\mathrm{P}(S) &= \frac{n-k}{n}\cdot\frac{k}{n+k} + \frac{k}{n}\cdot\frac{k(k-1)+n(n-1)}{(n+k)(n+k-1)} \\ &= \frac{k\big[(n-k)(n+k-1)+k^2-k+n^2-n\big]}{n(n+k)(n+k-1)}.\end{aligned}$$

대괄호 안에서 $(n-k)(n+k-1) = n^2-k^2-n+k$이므로, 대괄호 전체는 $2n^2-2n = 2n(n-1)$이다. 그러므로

$$\boxed{\mathrm{P}(S) = \frac{2k(n-1)}{(n+k)(n+k-1)}}$$

최댓값 구하기. $k=0$이면 확률은 0이므로 $1\le k\le n$으로 두고 역수를 최소화한다. 양의 상수 $2(n-1)$을 제외하면

$$\frac{(n+k)(n+k-1)}{k} = k + (2n-1) + \frac{n(n-1)}{k},$$

이므로 정수 범위에서 $g(k) = k + \dfrac{n(n-1)}{k}$를 최소화하면 된다. 이웃한 값끼리 비교하면

$$g(k+1) - g(k) = 1 - \frac{n(n-1)}{k(k+1)}.$$

이 값은 $k(k+1) < n(n-1)$일 때, 즉 $k < n-1$일 때 음수이고, $k = n-1$이면 $k(k+1) = (n-1)n$이므로 정확히 0이다. 따라서 g는 $k=1$부터 $k=n-1$까지 순감소하고, $g(n) = g(n-1)$이다. 그러므로 확률이 최대가 되는 것은

$$k = n-1 \quad \text{그리고} \quad k = n,$$

일 때이며, $n\ge 2$이므로 둘 다 허용된 범위 안에 있다. $k=n-1$일 때 최댓값은

$$\frac{2(n-1)^2}{(2n-1)(2n-2)} = \frac{n-1}{2n-1},$$

이고, $k=n$일 때도 $\dfrac{2n(n-1)}{2n(2n-1)} = \dfrac{n-1}{2n-1}$로 같은 값을 준다. ($g$를 연속함수로 볼 때의 최소점은 $\sqrt{n(n-1)}$이며, 이는 $n-1$과 n 사이에 엄격하게 놓인다. 두 정수에서 값이 같다는 것은 근사가 아니라 정확한 사실이다.)

검산

- 소문항 (i)에서 $k=n$이면 $\frac{1}{2}$이다. 이때 $2n$개의 말이 모두 Q를 거치고 그중 n개가 돌아오므로, 검은 말은 어느 주머니에서 끝날 가능성도 똑같다. 맞다.
- 소문항 (ii)에서 $k=0$이면 0이 나오는데, 검은 말들이 한 번도 만나지 않으므로 옳다. $n=2, k=1$이면 $\frac{2}{3\cdot 2} = \frac{1}{3}$이 나오며, 이는 손으로 직접 확인할 수 있다.

- $2 \le n \le 7$인 경우와 모든 k에 대하여 가능한 모든 추출 쌍을 (분수로) 정확히 나열해 보았다. 두 공식 모두 모든 경우에 일치했고, (ii)의 최대점은 항상 $\{n-1, n\}$이었다. 예를 들어 $n = 4$일 때 최댓값은 $\frac{3}{7} \approx 0.4286$이다.

흔한 실수

- 두 번째 추출에서 $k/(n+k)$ 대신 k/n을 쓰는 실수: 주머니 Q에는 말이 $n+k$개로 늘어나 있다.
- 소문항 (ii)의 경우 2에서 "둘 다 뽑히지 않는" 가능성을 빠뜨리는 실수: 두 검은 말이 모두 Q에 남아도 여전히 함께 있다.
- 미적분으로 최댓값을 구한 뒤 $\sqrt{n(n-1)}$을 정수 하나로 반올림하는 실수. 이산적인 비교를 하면 최대점이 두 개이고 그 값이 같음을 알 수 있다.

STEP 너머

두 소문항의 바탕에 있는 도구는 **초기하분포**(hypergeometric distribution)이다. 이는 비복원추출(without replacement)로 뽑은 무작위 표본 안에 표시된 원소가 몇 개 있는지를 나타내는 분포이다. 첫 단계의 결과에 대해 조건을 걸고(conditioning) 그것을 두 번째 단계에 이용하는 것이 바로 전확률 법칙이며, 이와 같은 두 단계 추론은 에렌페스트(Ehrenfest) 항아리 모형처럼 용기 사이의 섞임과 확산을 다루는 모형의 바탕이 된다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 초기하분포와 조건부 확률을 바로 이런 방식으로 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.