풀이

STEP II 1992, 문제 16

확률과 통계 • 조건부 확률

> **STEP II 1992, 16번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/92-s2-q16

이 문제는 *Beyond STEP*의 12.17 절 **바나흐의 성냥갑**에서 단계별 안내와 함께 자세히 다룬다:
https://beyondstep.keithmcnulty.org/ko/ch12/matchboxes.html

핵심 아이디어

이 문제는 바나흐(Banach)의 성냥갑 문제(matchbox problem)의 한 변형이다. 부족함을 언제 발견했는지(시도 N에서) 주어져 있으므로, 이것은 조건부 확률(conditional probability)이다. 따라서 두 개의 결합 확률(joint probability)을 계산한다. 하나는 첫 실패가 시도 N에서 일어나고 그것이 토피일 확률이고, 다른 하나는 첫 실패가 시도 N에서 일어나고 그것이 박하사탕일 확률이다. 각각은 "처음 $N-1$번의 시도에 각 종류가 정확히 몇 번씩 들어 있고, 그다음에 특정한 한 종류가 나온다"는 계산이며, 이는 이항(binomial) 계산이다.

풀이

각 시도는 독립(independent)이고, 매번 확률 p로 토피를, 확률 q로 박하사탕을 집으려 한다고 가정한다. 단, $0 < p < 1$이다. T를 N번째 시도에서 그녀가 토피를 집으려다 토피 봉지가 비어 있음을 발견하고, 그 이전 시도에서는 빈 봉지를 발견한 적이 없는 사건(event)이라 하자. M을 토피 대신 박하사탕으로 바꾼 같은 사건이라 하자. 구하려는 것은

$$\mathrm{P}(\text{토피} \mid \text{첫 발견 시점이 시도} N) = \frac{\mathrm{P}(T)}{\mathrm{P}(T) + \mathrm{P}(M)}.$$

이다.

T의 확률

T가 일어나려면 $N-1$번의 시도 후에 토피 봉지가 비어 있어야 한다. 지금까지 실패한 시도가 없으므로, 그 시도들은 모두 사탕 하나씩을 꺼냈다. 따라서 처음 $N-1$번의 시도 중 정확히 n번은 토피를 집으려 한 것이고(한 번 더 있었다면 실패였을 것이다), 나머지 $N-1-n$번은 박하사탕을 집으려 한 것이다. 그리고 시도 N은 토피를 집으려 한 것이다.

역으로, 처음 $N-1$번의 시도에 토피 시도가 정확히 n번, 박하사탕 시도가 $N-1-n \leq m$번 순서에 관계없이 들어 있다고 하자. 그러면 이전의 어떤 시도도 실패했을 수 없다. 어느 단계에서도 그녀는 n개보다 많은 토피나 m개보다 많은 박하사탕을 집으려 한 적이 없기 때문이다. 이때 시도 N이 토피이면 T가 일어난다.

> **?** **생각해 보기** 처음 $N-1$번의 시도의 순서가 중요하지 않을까? 일찍 다 떨어지는 일이 없도록 말이다.

그렇지 않다. 그녀는 봉지가 빈 상태에서 그 봉지에 손을 넣을 때에만 봉지가 비었음을 **발견한다**. 토피 봉지를 일찍(즉 n번째 토피에서) 비우더라도, 시도 N 전에 다시 토피를 집으려 하지만 않으면 괜찮다. 처음 $N-1$번의 시도에는 토피 시도가 n번뿐이므로 이것은 저절로 성립한다. 따라서 $\binom{N-1}{n}$가지 순서가 모두 허용되고,

$$\mathrm{P}(T) = \binom{N-1}{n} p^n q^{N-1-n} \cdot p = \binom{N-1}{n} p^{n+1} q^{N-1-n}, \qquad n \leq N-1 \leq n+m,$$

이며, 그 밖의 경우에는 $\mathrm{P}(T) = 0$이다. $N-1 < n$이면 토피를 다 써 버렸을 수 없다. $N-1 > n+m$이면 m개보다 많은 박하사탕이 필요했을 것이다.

M의 확률

역할을 바꾸어 똑같이 하면

$$\mathrm{P}(M) = \binom{N-1}{m} q^{m+1} p^{N-1-m}, \qquad m \leq N-1 \leq n+m,$$

이며, 그 밖의 경우에는 $\mathrm{P}(M) = 0$이다.

답

발견은 시도 $\min(n,m)+1$부터 $n+m+1$ 사이 어딘가에서 일어나야 한다. 시도 $n+m+2$는 불가능하다. 그때쯤이면 두 봉지가 모두 비어 있을 것이므로 이미 어떤 시도가 실패했을 것이기 때문이다.

- $N \leq m$이면($\mathrm{P}(M) = 0$이지만 $\mathrm{P}(T) > 0$이므로) 그녀는 확실히 토피를 집으려 한 것이다. 확률은 1이다.
- $N \leq n$이면($\mathrm{P}(T) = 0$이므로) 확률은 0이다.

- $\max(n,m)+1 \le N \le n+m+1$이면 두 경우가 모두 가능하고,

$$\mathrm{P}(\text{토피}) = \frac{\binom{N-1}{n} p^{n+1} q^{N-1-n}}{\binom{N-1}{n} p^{n+1} q^{N-1-n} + \binom{N-1}{m} q^{m+1} p^{N-1-m}}.$$

  이다.

마지막 경우에는 분자와 분모를 $\mathrm{P}(T)$로 나누어 식을 정리할 수 있다. 이항계수(binomial coefficient)에서 공통인수 $(N-1)!$이 약분되고, 거듭제곱들은 $(q/p)^{n+m+2-N}$으로 합쳐진다. 그러면

$$\mathrm{P}(\text{토피}) = \left[1 + \frac{n!\,(N-1-n)!}{m!\,(N-1-m)!} \left(\frac{q}{p}\right)^{n+m+2-N}\right]^{-1}.$$

을 얻는다.

검산

- 대칭성(symmetry): $n=m$이고 $p=q=\frac{1}{2}$이면 두 식이 같으므로 답은 당연히 $\frac{1}{2}$이다.
- $n=2$, $m=3$, $p=\frac{1}{2}$, $N=5$로 놓자. 그러면 $\mathrm{P}(T) = \binom{4}{2}/2^5 = \frac{3}{16}$이고 $\mathrm{P}(M) = \binom{4}{3}/2^5 = \frac{1}{8}$이므로, 답은 $\frac{3/16}{5/16} = \frac{3}{5}$이다. 토피가 더 적으므로 토피가 떨어질 가능성이 더 크다.
- 결국에는 어떤 시도가 반드시 실패하므로 $N = 1, \ldots, n+m+1$에 대하여 $\sum_N \big(\mathrm{P}(T) + \mathrm{P}(M)\big) = 1$이다. 이 사실과 공식을 모든 $n, m \le 6$와 여러 p 값에 대하여 한 단계씩 정확히 열거한 결과와 대조하여 확인하였다. $n=4$, $m=3$, $N=7$, $p=0.55$로 2×10^5회 시뮬레이션(simulation)한 결과는 0.527로, 정확한 값 0.528과 비교된다.

흔한 실수

- N번째 시도 자체에 해당하는 마지막 인수 p(또는 q)를 빠뜨린다.
- 조건부 확률 $\mathrm{P}(T)/(\mathrm{P}(T)+\mathrm{P}(M))$ 대신 $\mathrm{P}(T)$를 답으로 쓴다. 발견이 시도 N에서 일어났다는 것이 주어져 있다.
- 사건을 "토피가 시도 N에서 떨어진다"로 생각한다. 마지막 토피를 꺼내는 것은 발견이 아니다. 발견은 그다음 시도에서 일어난다.

STEP 너머

첫 실패의 시점은 두 **음이항** 분포(negative binomial distribution)의 혼합(mixture)이다. 음이항 분포는 주어진 횟수의 성공을 얻는 데 필요한 시행 횟수를 센다. 바나흐의 원래 성냥갑 문제($n = m, p = \frac{1}{2}$)는 발견하는 순간 다른 상자에 성냥이 몇 개 남아 있는지를 묻는다. 그 답은 스털링 공식(Stirling's formula)으로 근사할 수 있으며, 평균(mean)은 약 $2\sqrt{n/\pi}$이다. 음이항 분포에 대해서는 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)을 참고하라. 펠러(Feller)의 고전 *An Introduction to Probability Theory and Its Applications*(확률론과 그 응용 입문) 제1권에는 성냥갑 문제에 대한 유명한 논의가 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.