풀이

STEP I 2006, 문제 14

확률과 통계 • 비복원추출

> **STEP I 2006, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/06-s1-q14

핵심 아이디어

r번째 시도에서 빨간 사탕을 꺼낸다는 것은 파란 사탕을 $r-1$번 꺼낸 뒤 빨간 사탕을 꺼낸다는 뜻이다. 복원추출(with replacement)에서는 매번의 추출이 똑같은 상황이므로 확률은 거듭제곱의 단순한 곱이 된다. 그다음 n을 연속변수로 보고 로그(logarithm)를 미분하여 n에 대해 최댓값을 구한다. 비복원추출(without replacement)에서는 사탕을 **모두** 꺼낸다고 상상하는 것이 가장 깔끔하다. 그러면 빨간 사탕은 어느 위치에 있을 가능성이나 똑같다.

풀이

이 풀이 전체에서 $n \geq 0$은 정수이고, R는 빨간 사탕을 처음 꺼내는 시도의 번호이다.

소문항 (i)

꺼낸 사탕을 다시 넣고 주머니를 흔들기 때문에 각 추출은 서로 독립(independent)이며, 매번 빨간 사탕은 확률 $\frac{1}{n+1}$로, 파란 사탕은 확률 $\frac{n}{n+1}$로 꺼내진다. 사건 $R=r$는 처음 $r-1$번은 파란 사탕이고 r번째는 빨간 사탕이라는 뜻이므로, 독립성에 의해

$$P(R=r) = \left(\frac{n}{n+1}\right)^{r-1} \frac{1}{n+1} = \frac{n^{r-1}}{(n+1)^r}, \qquad r = 1, 2, 3, \ldots$$

이다. (이것은 성공 확률이 $\frac{1}{n+1}$인 기하분포(geometric distribution)이다.)

이제 r를 고정하고 이를 n의 함수로 생각하자.

$$f(n) = \frac{n^{r-1}}{(n+1)^r}.$$

$r=1$이면 $f(n) = \frac{1}{n+1}$이고, 이는 $n=0$에서 가장 크다(파란 사탕이 없으면 첫 사탕은 반드

시 빨간색이다). 따라서 $r \geq 2$라고 하자. 그러면 $f(0) = 0$이고 $n > 0$에서 $f(n) > 0$이므로, n을 실수 변수로 보고 $n > 0$에서 로그를 취할 수 있다.

$$\ln f(n) = (r-1)\ln n - r\ln(n+1), \qquad \frac{\mathrm{d}}{\mathrm{d}n}\ln f(n) = \frac{r-1}{n} - \frac{r}{n+1} = \frac{(r-1)-n}{n(n+1)}.$$

? **생각해 보기** n은 자연수여야 한다. 미분해도 괜찮은가?

괜찮다. 도함수(derivative)는 f가 어디에서 증가하고 감소하는지 알아내는 데에만 쓰기 때문이다. $n > 0$에서 분모 $n(n+1)$은 양수이므로, 도함수는 $0 < n < r-1$에서 양수이고 $n > r-1$에서 음수이다. 따라서 f는 $[0, r-1]$에서 순증가하고 $[r-1, \infty)$에서 순감소한다. $r-1$ 자체가 정수이므로 이 값이 다른 모든 정수보다 크며, 이웃한 정수끼리 비교할 필요가 없다.

그러므로 모든 경우($r = 1$ 포함)에 확률은

$$n = r - 1,$$

일 때 최대가 되고, 최댓값은 $\dfrac{(r-1)^{r-1}}{r^r}$이다($r = 1$일 때는 1로 읽는다). 예를 들어 $r = 3$이면 가장 좋은 주머니에는 파란 사탕이 $n = 2$개 들어 있고, 확률은 $\frac{4}{27}$이다.

소문항 (ii)

이제는 추출이 독립이 아니지만, 쉬운 방법이 두 가지 있다.

직접 곱하기. $R = r$이려면 처음 $r-1$번 모두 파란 사탕이 나오고 그다음 빨간 사탕이 나와야 한다. 이는 파란 사탕이 충분히 많을 때, 즉 $r - 1 \leq n$일 때에만 가능하다. 앞선 추출이 모두 파란색이었다는 조건 아래에서, k번째 추출은 빨간 사탕 하나를 포함한 $n + 2 - k$개의 사탕이 든 주머니에서 이루어진다. 따라서

$$P(R = r) = \frac{n}{n+1} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdots \frac{n-r+2}{n-r+3} \cdot \frac{1}{n-r+2}.$$

각 분자가 다음 분모와 약분되어, 이 곱은 연쇄적으로 소거되어(telescoping)

$$P(R = r) = \begin{cases} \dfrac{1}{n+1}, & 1 \leq r \leq n+1, \\ 0, & r > n+1. \end{cases}$$

가 된다.

대칭성. 빨간 사탕에서 멈추든 멈추지 않든 빨간 사탕이 차지하는 위치는 달라지지 않는다. 그러니 주머니가 빌 때까지 계속 꺼낸다고 상상하자. (서로 구별되는) 사탕들의 $(n+1)!$가지 순서는 모두 같은 확률로 나타나므로, 빨간 사탕은 $n+1$개의 위치 중 어디에 있을 가능성이나 똑같다. 이로부터 같은 답을 얻는다.

최댓값을 구하기 위해 r를 고정하자. $n < r-1$이면 확률은 0이다. $n \geq r-1$이면 확률은 $\frac{1}{n+1}$이고, 이는 n이 커질수록 순감소한다. 따라서 확률은 허용되는 가장 작은 주머니

$$n = r - 1,$$

에서 최대가 되며, 최대 확률은 $\frac{1}{r}$이다. ($n = r-1$이면 빨간 사탕은 r개의 사탕 중 어느 것일 가능성이나 똑같고, 그것을 마지막에 꺼내는 것은 같은 확률을 가진 r가지 결과 중 하나이다.)

검산

- (i)에서 $\sum_{r\geq 1} P(R=r) = \frac{1}{n+1} \cdot \dfrac{1}{1-n/(n+1)} = 1$이고, (ii)에서 $n+1$개의 같은 값 $\frac{1}{n+1}$을 더하면 1이다.
- (i)에서 $r=3$일 때: $n=1,2,3$은 각각 $\frac{1}{8}$, $\frac{4}{27} \approx 0.148$, $\frac{9}{64} \approx 0.141$을 주므로 $n=2$가 가장 크다.
- $n \leq 6$에 대해 사탕의 모든 순서를 나열하여 (ii)를 확인하였다. $n=3, r=2$로 (i)을 모의실험한 결과는 0.188이었고, 정확한 값은 $\frac{3}{16} = 0.1875$이다.

흔한 실수

- (i)에서 $\left(\frac{n}{n+1}\right)^r \frac{1}{n+1}$이라고 쓰는 것: 빨간 사탕 앞에 오는 파란 사탕은 $r-1$개뿐이다.
- (ii)에서 $r > n+1$일 때 답이 0이라는 것을 잊는 것. 바로 이 점 때문에 "n을 가능한 한 작게 하라"는 결론이 $n=0$까지 내달리지 않는다.

STEP 너머

두 최댓값 문제는 사실 **최대가능도 추정**(*maximum likelihood estimation*)이다. 빨간 사탕이 r번째 시도에서 처음 나왔다는 것은 보았지만 n은 모른다면, 관찰한 결과가 나올 확률을 가장 크게 하는 n의 값은 두 실험 모두에서 $n = r-1$이다. 이것은 대학 1학년 통계학의 핵심 아이디어이다. (ii)의 대칭성 논증은 **교환가능성**(*exchangeability*)의 한 예로, 이 성질 덕분에 많은 비복원추출 문제가 쉬워진다. 두 아이디어 모두 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에서 자세히 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.