풀이

STEP I 2010, 문제 12

확률과 통계 • 꼬리 합을 통한 기댓값

> **STEP I 2010, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/10-s1-q12

핵심 아이디어

꼬리 합 공식(tail-sum formula)은 $n\,\mathrm{P}(X=n)$을 $\mathrm{P}(X=n)$을 n번 더한 것으로 쓴 다음, 항들을 다른 순서로 모으는 데서 나온다. 펭귄 문제에서는 사건 $X \geq n$을 $X = n$보다 훨씬 쉽게 기술할 수 있다. $n-1$개의 상자를 연 뒤에도 상자를 더 열어야 하는 것은 정확히 그 $n-1$개의 상자에 모두 같은 종류의 펭귄이 들어 있을 때이다. 따라서 꼬리 확률은 간단하고, 방금 증명한 공식을 쓰면 $\mathrm{E}(X)$는 두 개의 등비급수(geometric series)가 된다.

풀이

정의와 꼬리 합 공식

X는 $1, 2, 3, \ldots$의 값만 가지므로, 다음 급수가 수렴(converge)한다면

$$\mathrm{E}(X) = \sum_{m=1}^{\infty} m\,\mathrm{P}(X=m),$$

이다.

각 $n \geq 1$에 대하여 사건 $X \geq n$은 $m \geq n$인 사건 $X = m$들의 서로소 합집합(disjoint union)이므로 $\mathrm{P}(X \geq n) = \sum_{m \geq n} \mathrm{P}(X=m)$이다. 따라서

$$\begin{aligned}\sum_{n=1}^{\infty} \mathrm{P}(X \geq n) &= \sum_{n=1}^{\infty} \sum_{m=n}^{\infty} \mathrm{P}(X=m) \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{m} \mathrm{P}(X=m) \\ &= \sum_{m=1}^{\infty} m\,\mathrm{P}(X=m) = \mathrm{E}(X).\end{aligned}$$

? **생각해 보기** 두 합의 순서를 바꾸어도 되는 이유는 무엇일까?

이중합은 $1 \le n \le m$인 모든 쌍 (n, m)에 대하여 $\mathrm{P}(X = m)$을 더한다. 첫째 줄에서는 n을 고정하고 m에 대해 더하고, 둘째 줄에서는 m을 고정하고 $n = 1, \ldots, m$의 m개 값에 대해 더한다. 모든 항이 음이 아니며, 음이 아닌 항들로 이루어진 급수의 합은 더하는 순서에 의존하지 않는다(양변이 무한대일 수도 있는데, 그런 경우에는 양변 모두 무한대이다). 항들을 삼각형 모양으로 늘어놓고 m번째 행에 $\mathrm{P}(X = m)$을 m개 놓는다고 생각하면 도움이 된다. 행별로 더하면 $\mathrm{E}(X)$가 되고, 열별로 더하면 꼬리 합이 된다. 이로써 공식이 증명된다.

꼬리 확률

문제의 의도대로 서로 다른 상자의 내용물은 독립(independent)이라고 가정한다. D_k를 상자 k에 아빠 펭귄이 들어 있는 사건이라 하면 $\mathrm{P}(D_k) = p, \mathrm{P}(D_k') = q$이다.

$n \ge 2$라 하자. $X \ge n$인 것은 정확히 처음 $n-1$개의 상자에 아직 두 종류가 하나씩 **모이지 않았을** 때, 즉 모두 아빠 펭귄이거나 모두 엄마 펭귄일 때이다. 이 두 경우는 서로소이므로 독립성에 의해

$$\mathrm{P}(X \ge n) = p^{n-1} + q^{n-1} \qquad (n \ge 2).$$

특히 $n = 4$이면

$$\mathrm{P}(X \ge 4) = p^3 + q^3,$$

이 되어 원하는 결과를 얻는다. 또한 한 상자에 두 종류가 모두 들어 있을 수는 없으므로 항상 $X \ge 2$이고, 따라서 $\mathrm{P}(X \ge 1) = \mathrm{P}(X \ge 2) = 1$이다. (위의 식은 $n = 2$일 때 $p + q = 1$을 주므로 일치하지만, $n = 1$일 때는 2를 주므로 $n = 1$인 항은 따로 다루어야 한다.)

기댓값

꼬리 합 공식에 의해

$$\begin{aligned}\mathrm{E}(X) &= \mathrm{P}(X \ge 1) + \sum_{n=2}^{\infty} (p^{n-1} + q^{n-1}) \\ &= 1 + \sum_{k=1}^{\infty} p^k + \sum_{k=1}^{\infty} q^k \\ &= 1 + \frac{p}{1-p} + \frac{q}{1-q},\end{aligned}$$

이다. 여기서 $0 < p < 1, 0 < q < 1$이므로 두 등비급수는 모두 수렴한다($p \ne 0, q \ne 0$이라는 조건은 바로 여기서 쓰인다). $1 - p = q, 1 - q = p$이므로

$$\mathrm{E}(X) = 1 + \frac{p}{q} + \frac{q}{p} = 1 + \frac{p^2 + q^2}{pq}.$$

이제 $p^2+q^2=(p+q)^2-2pq=1-2pq$이므로

$$\mathrm{E}(X)=1+\frac{1-2pq}{pq}=1+\frac{1}{pq}-2=\frac{1}{pq}-1,$$

이 되어 원하는 결과를 얻는다.

마지막 부분

양수 p와 q에 대한 산술-기하 평균 부등식(AM–GM inequality)에 의해

$$\sqrt{pq}\le\frac{p+q}{2}=\frac{1}{2},\qquad \text{이므로}\qquad 0<pq\le\frac{1}{4}.$$

? **생각해 보기** 역수를 취하면 부등호의 방향은 어떻게 될까?

$pq>0$이므로 역수를 취하면 부등호의 방향이 바뀐다. 즉 $\frac{1}{pq}\ge 4$이다. 따라서

$$\mathrm{E}(X)=\frac{1}{pq}-1\ge 4-1=3,$$

이고, 등호는 정확히 $p=q=\frac{1}{2}$일 때 성립한다. (같은 이야기로, $pq=p(1-p)=\frac{1}{4}-(p-\frac{1}{2})^2\le\frac{1}{4}$이다.)

검산

- $p=q=\frac{1}{2}$일 때: $X-1$은 상자 1과 다른 첫 상자가 나올 때까지 기다리는 횟수로, 평균이 2인 기하 확률변수이다. 따라서 $\mathrm{E}(X)=3$이고, 이는 $\frac{1}{1/4}-1$과 일치한다.
- $p\to 0$이면 $\mathrm{E}(X)\to\infty$이다. 아빠 펭귄이 드물어서 오래 기다려야 하기 때문이다. 이는 $\frac{1}{pq}-1$이 한없이 커지는 것과 일치한다.
- 직접 계산하면 $n\ge 2$일 때 $\mathrm{P}(X=n)=p^{n-1}q+q^{n-1}p$이다. $p=\frac{1}{3}$에 대하여 $n\,\mathrm{P}(X=n)$을 수치적으로 더하면 3.5가 나오고, $p=0.3$으로 수집가 200 000명을 모의실험하면 평균 3.761이 나와 정확한 값 $\frac{1}{0.21}-1\approx 3.762$와 비교된다.

흔한 실수

- $n=1$에도 $p^{n-1}+q^{n-1}$을 사용하는 것. 그러면 "확률"이 2가 되고 기댓값이 1만큼 커진다.
- $\mathrm{P}(X\ge 4)=p^4+q^4$라고 쓰는 것. 네 번째 상자가 필요하려면 처음 세 개의 상자만 같으면 된다.

- 마지막 부분에서 양수 사이의 부등식에서 역수를 취하면 부등호의 방향이 바뀐다는 것을 잊는 것.

STEP 너머

꼬리 합 공식에는 연속형 쌍둥이가 있다. 음이 아닌 확률변수(random variable)에 대하여 $\mathrm{E}(X) = \int_0^\infty \mathrm{P}(X > x)\,\mathrm{d}x$이며, 이것도 정확히 같은 방식으로 순서를 바꾸어 증명한다(일반적인 경우에는 토넬리 정리(Tonelli's theorem)가 이를 정당화한다). 펭귄 문제는 **쿠폰 수집가 문제**(*coupon collector's problem*)에서 종류가 두 가지인 경우이다. 똑같이 나올 법한 종류가 N가지이면 필요한 상자 수의 기댓값은 $N(1 + \frac{1}{2} + \cdots + \frac{1}{N})$으로, 대략 $N \ln N$이다. 두 아이디어 모두 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)과 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에 나온다.

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.