풀이

STEP II 1991, 문제 15

확률과 통계 • 기댓값의 선형성

> **STEP II 1991, 15번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/91-s2-q15

핵심 아이디어

순서쌍 (n_1, n_2)를 $m \times m$ 격자의 한 칸으로 생각하자. m^2개의 칸은 모두 같은 확률로 나온다. $|n_1 - n_2| = k$인 칸들은 주대각선과 평행한 두 대각선 위에 놓이므로 세기 쉽다. 그다음은 모두 **기댓값의 선형성**(*linearity of expectation*)에서 나온다. 이웃한 항들이 변수를 공유하여 독립(independent)이 아니더라도, 차들의 합의 기댓값은 각 기댓값의 합과 같다. 감독관이 걷는 거리도 같은 종류의 합이다. 다만 첫 구간은 무작위 책상이 아니라 정해진 책상에서 출발한다는 점이 다르다.

풀이

$|n_1 - n_2|$의 분포

n_1과 n_2는 서로 독립이고 각각 $\{1, \ldots, m\}$ 위에서 균등분포(uniform distribution)를 따르므로, m^2개의 순서쌍 (n_1, n_2)는 각각 확률 $1/m^2$를 가진다.

$1 \le k \le m-1$인 k를 고정하자. $n_1 - n_2 = k$인 순서쌍은 $(k+1, 1), (k+2, 2), \ldots, (m, m-k)$로 $m-k$개이다. 대칭성에 의해 $n_2 - n_1 = k$인 순서쌍도 $m-k$개 있고, $k \neq 0$이므로 두 집합은 겹치지 않는다. 따라서

$$\mathrm{P}(|n_1 - n_2| = k) = \frac{2(m-k)}{m^2},$$

이고, 이것이 구하려던 결과이다. ($k = 0$일 때는 순서쌍이 m개이므로 $\mathrm{P}(|n_1 - n_2| = 0) = 1/m$이다. 검산하면 $\frac{1}{m} + \sum_{k=1}^{m-1} \frac{2(m-k)}{m^2} = \frac{1}{m} + \frac{m(m-1)}{m^2} = 1$이다.)

$|n_1 - n_2|$의 기댓값

$k = 0$인 값은 기댓값(expected value)에 아무것도 더하지 않으므로

$$\begin{aligned}\mathrm{E}|n_1 - n_2| &= \sum_{k=1}^{m-1} k \cdot \frac{2(m-k)}{m^2} = \frac{2}{m^2}\left(m\sum_{k=1}^{m-1} k - \sum_{k=1}^{m-1} k^2\right) \\ &= \frac{2}{m^2}\left(\frac{m \cdot m(m-1)}{2} - \frac{(m-1)m(2m-1)}{6}\right) \\ &= \frac{2}{m^2} \cdot \frac{m(m-1)(3m-(2m-1))}{6} = \frac{(m-1)(m+1)}{3m} = \frac{m^2-1}{3m},\end{aligned}$$

이다. 여기서 $n = m-1$로 두고 잘 알려진 합 공식 $\sum_{k=1}^{n} k = \frac{1}{2}n(n+1)$과 $\sum_{k=1}^{n} k^2 = \frac{1}{6}n(n+1)(2n+1)$을 사용하였다.

$m = 2$이고 정수가 세 개인 경우

$m = 2$이면 같은 확률로 나오는 세 수의 순서쌍 (n_1, n_2, n_3)이 $2^3 = 8$개 있다. $S = |n_1 - n_2| + |n_2 - n_3|$라 하면 다음과 같다.

(n_1, n_2, n_3)	111	112	121	122	211	212	221	222
S	0	1	2	1	1	2	1	0

합계가 8이므로 $\mathrm{E}(S) = 8/8 = 1$이다. 주어진 값은 $2(m^2-1)/(3m) = 2 \cdot 3/6 = 1$이므로 일치한다.

> **생각해 보기** $|n_1 - n_2|$와 $|n_2 - n_3|$는 둘 다 n_2를 포함한다. 그래도 그냥 기댓값을 더해도 될까?

그렇다. 기댓값의 선형성 $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$에는 독립성이 전혀 필요하지 않다. 각 순서쌍 (n_i, n_{i+1})만 따로 보면 서로 독립인 두 균등분포 정수로 이루어져 있으므로, 각 항의 기댓값은 $(m^2-1)/(3m)$이다. 항이 $r-1$개이므로

$$\mathrm{E}(|n_1 - n_2| + |n_2 - n_3| + \cdots + |n_{r-1} - n_r|) = \frac{(r-1)(m^2-1)}{3m}.$$

$r = 3$인 경우가 바로 앞에서 $m = 2$에 대해 확인한 결과이다.

감독관

줄을 따라 책상에 $1, 2, \ldots, m$의 번호를 붙이면, 책상 i와 j 사이의 거리는 $d|i-j|$이다. $j = 1, \ldots, N+1$에 대하여 j번째로 부른 책상의 번호를 c_j라 하자. 부르는 사람들은 서로 독립이고 각각 $\{1, \ldots, m\}$ 위에서 균등분포를 따른다고 모형화한다. 이것이 "m명의 수험생 중 누구에게나 같은 확률로 불린다"를 해석한 방식이다. 수험생들은 동시에 부르거나 다른 호출 중에 부르지 않으므로, 감독관은 호출을 하나씩 차례로 처리하며 $j+1$번째 호출이 들

어올 때 책상 c_j에 서 있다. (같은 수험생이 연달아 두 번 부르면 그 구간의 길이는 0이며, 공식은 이 경우도 포함한다.)

따라서 걸은 거리의 합은

$$D = d\,|c_1 - 1| + d\sum_{j=1}^{N} |c_{j+1} - c_j|.$$

첫 구간. 감독관은 책상 1에서 출발하고 c_1은 균등분포를 따르므로

$$\mathrm{E}|c_1 - 1| = \frac{1}{m}\sum_{i=1}^{m}(i-1) = \frac{m-1}{2}.$$

이 값이 $(m^2-1)/(3m)$보다 크다는 점에 주목하자. 줄의 끝에서 출발하기 때문에 첫 번째 걸음이 평균보다 길어진다.

나머지 N개의 구간. 여기서는 $c_1, \ldots, c_{N+1}$이 $r = N+1$일 때의 $n_1, \ldots, n_r$ 역할을 하므로, 앞의 결과에서

$$\mathrm{E}\sum_{j=1}^{N} |c_{j+1} - c_j| = \frac{N(m^2-1)}{3m}.$$

합치기. 다시 선형성에 의해

$$\begin{aligned}\mathrm{E}(D) &= d\left(\frac{m-1}{2} + \frac{N(m-1)(m+1)}{3m}\right)\\ &= \frac{d(m-1)}{6m}(3m + 2N(m+1)) = \frac{d(m-1)}{6m}[2N(m+1) + 3m],\end{aligned}$$

이고, 이것이 구하려던 결과이다.

검산

- $m = 1$이면 책상이 하나뿐이므로 감독관은 전혀 움직이지 않는다. 공식도 마땅히 0을 준다.
- $N = 0$이면 첫 호출만 있고, 공식은 첫 구간의 기댓값인 $d(m-1)\cdot 3m/(6m) = \frac{1}{2}d(m-1)$로 줄어든다.
- 부르는 사람의 가능한 모든 순서를 정확히 열거하면 $m \le 4$, $N \le 3$인 모든 경우에 공식이 맞음을 확인할 수 있다. $m = 10$, $N = 20$, $d = 1$로 2×10^5회 시뮬레이션한 결과 평균은 70.50이었고, 공식의 값은 $\frac{9}{60}(440 + 30) = 70.5$이다.

흔한 실수

- 첫 구간을 다른 구간과 똑같이 취급하는 것. 감독관은 무작위 책상이 아니라 책상 1에서 출발하므로, 그 구간의 기댓값은 $(m^2-1)/(3m)$이 아니라 $\frac{1}{2}(m-1)$이다.

- $n_1 - n_2 = k$인 경우만 세고 $n_2 - n_1 = k$에 해당하는 인수 2를 빠뜨리거나, $k = 0$인 경우를 잘못 두 배로 세는 것.
- 합의 항들이 서로 독립이어야만 기댓값을 더할 수 있다고 생각하는 것.

STEP 너머

감독관의 위치는 **마르코프 연쇄**(*Markov chain*)이다. 다음에 어디로 갈지는 오직 지금 어디에 있는지에만 의존한다. 여기서는 연쇄가 매우 단순해서 각 단계가 과거를 완전히 "잊는다". 더 풍부한 모형(예를 들어 가까운 책상에서 호출할 가능성이 더 큰 경우)에서는 호출당 장기 평균 이동 거리를 구하기 위해 연쇄의 정상 분포(stationary distribution)가 필요하다. 첫 번째 결과의 연속 버전은 $[0,1]$ 위의 서로 독립인 두 균등분포 점 사이의 평균 거리가 $\frac{1}{3}$이라는 것이며, 이는 $(m^2-1)/(3m)$을 m으로 나눈 값의 극한이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문서)은 기댓값의 선형성과 마르코프 연쇄를 모두 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.