풀이

STEP I 1999, 문제 13

확률과 통계 • 지시변수

> **STEP I 1999, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/99-s1-q13

핵심 아이디어

단위의 개수를 직접 세지 말고, 대신 **틈**을 센다. 단위의 개수는 이웃한 자석이 서로 떨어져 있는 곳의 개수보다 하나 많다. 틈이 생기는지는 이웃한 두 자석이 같은 방향을 향하는지에만 달려 있고, 이 $N-1$개의 사건은 서로 독립(independent)인 공정한 동전 던지기임이 밝혀진다. 따라서 틈의 개수는 이항분포(binomial distribution)를 따르고, 기댓값(expectation)과 분산(variance)이 곧바로 나온다.

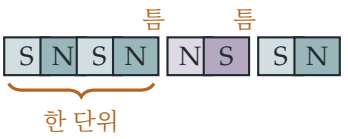
풀이

모형 세우기

"무작위로 놓는다"는 것을 각 자석이 다른 자석과 독립적으로 두 방향 중 어느 쪽이든 같은 확률로 향한다는 뜻으로 받아들인다. N극이 오른쪽에 있는 자석(SN)을 A**형**, N극이 왼쪽에 있는 자석(NS)을 B**형**이라 하자. 왼쪽부터 차례로 읽은 자석의 형을 $X_1, \ldots, X_N$이라 하자.

이웃한 두 자석을 보자. 서로 마주 보는 끝은 첫 번째 자석의 오른쪽 끝과 두 번째 자석의 왼쪽 끝이다.

- 같은 형(A 다음 A): 마주 보는 끝은 N과 S로 서로 반대이므로 **붙는다**. (B 다음 B도 S와 N으로 마찬가지이다.)
- 다른 형(A 다음 B): 마주 보는 끝은 N과 N이므로 **떨어진다**. (B 다음 A이면 S와 S가 된다.)



따라서 자석 i와 $i+1$ 사이에 틈이 생기는 것은 정확히 $X_i \neq X_{i+1}$일 때이다. $i = 1, \ldots, N-1$에 대하여 다음과 같이 정의한다.

$$G_i = \begin{cases} 1 & \text{만약} X_i \neq X_{i+1} \text{ (틈)}, \\ 0 & \text{만약} X_i = X_{i+1} \text{ (붙음)}. \end{cases}$$

틈이 하나 생길 때마다 줄이 한 조각씩 더 나뉘므로, 분리된 단위의 개수는 다음과 같다.

$$U = 1 + G_1 + G_2 + \cdots + G_{N-1}.$$

틈의 분포

각 G_i는 확률 $\frac{1}{2}$로 1이다. X_i가 무엇이든 X_{i+1}이 그와 다를 확률은 $\frac{1}{2}$이기 때문이다.

> **생각해 보기** 사건 "i에서의 틈"과 "$i+1$에서의 틈"은 둘 다 자석 $i+1$과 관련된다. G_i들은 정말 독립인가?

이웃한 G들이 자석을 공유하는데도 독립이다. 첫 번째 자석의 형 X_1을 고정하자. 그러면 수열 $G_1, \ldots, G_{N-1}$이 $X_2, X_3, \ldots, X_N$을 하나씩 차례로 결정하고($G_i = 0$이면 X_{i+1}은 X_i와 같고, $G_i = 1$이면 다른 형이다), 거꾸로 형들이 G들을 결정한다. 따라서 X_1이 어느 값이든, 같은 확률을 갖는 $(X_2, \ldots, X_N)$의 2^{N-1}가지 선택은 가능한 2^{N-1}가지 수열 $(G_1, \ldots, G_{N-1})$과 일대일로 대응한다. 그러므로 틈의 모든 수열은 확률 $1/2^{N-1}$을 가지며, 이는 바로 G_i들이 독립이고 각각 확률 $\frac{1}{2}$로 0 또는 1이라는 뜻이다.

따라서 $G_1 + \cdots + G_{N-1} \sim \mathrm{B}\left(N-1, \frac{1}{2}\right)$이다.

기댓값과 분산

$\mathrm{B}(n, p)$를 따르는 확률변수의 평균은 np, 분산은 $np(1-p)$이다. $n = N-1$, $p = \frac{1}{2}$로 두고, 상수 1을 더하면 평균은 이동하지만 분산은 변하지 않는다는 점을 기억하면 다음을 얻는다.

$$\mathrm{E}(U) = 1 + \tfrac{1}{2}(N-1) = \frac{N+1}{2},$$
$$\mathrm{Var}(U) = (N-1) \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \frac{N-1}{4}.$$

이항분포의 결과를 그대로 인용하고 싶지 않다면 이렇게 해도 된다. $\mathrm{E}(G_i) = \frac{1}{2}$이고, $G_i^2 = G_i$이므로 $\mathrm{Var}(G_i) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$이다. 선형성으로 평균을 얻고, 독립성 덕분에 $N-1$개의 분산을 더할 수 있다.

검산

- $N = 1$: 자석 하나는 언제나 한 단위이고, 공식도 평균 1, 분산 0을 준다.
- $N = 2$: 두 자석은 각각 확률 $\frac{1}{2}$로 1개 또는 2개의 단위를 이루므로 평균은 $\frac{3}{2}$이고 분산은 $\frac{5}{2} - \frac{9}{4} = \frac{1}{4}$로, 예측과 같다.
- $N \leq 12$에 대해 2^N가지 배열을 모두 정확히 열거하여 $\frac{N+1}{2}$과 $\frac{N-1}{4}$를 정확히 재현했고, $N = 20$으로 한 시뮬레이션에서는 평균 약 10.5, 분산 약 4.73을 얻었다(예측값은 10.5와 4.75).

흔한 실수

- **같은** 방향의 두 자석이 서로 밀어낸다고 생각하는 것. 두 자석은 반대 극(N과 S)으로 마주 보므로 서로 끌어당겨 붙는다.
- 독립성을 정당화하지 않고 분산을 더하는 것. 여기서 독립성은 참이지만 자명하지 않은데, 이웃한 틈들이 자석을 공유하기 때문이다.
- 틈을 $N - 1$개가 아니라 N개로 세어 $\mathrm{Var}(U) = \frac{N}{4}$로 쓰거나, 평균에서 추가되는 1을 빠뜨리는 것.

STEP 너머

단위 사이의 경계를 세어 단위를 세는 방법은 **지시변수**(indicator variable)의 방법으로, 확률론에서 가장 유용한 도구 가운데 하나이다. 개수를 0–1 변수들의 합으로 쓴 다음 기댓값의 선형성(linearity of expectation)을 이용하는데, 여기에는 독립성이 전혀 필요하지 않다. 이 문제에서 U는 두 기호로 이루어진 무작위 수열의 **연**(run)의 개수를 센다. 연의 개수는 통계학에서 수열이 무작위로 보이는지 검사하는 Wald–Wolfowitz 연 검정(runs test)의 바탕이 된다. 지시변수와 그 분산에 대해 더 알고 싶다면 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)를 보라.

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.