풀이

STEP III 2013, 문제 12

확률과 통계 • 지시 확률변수

> **STEP III 2013, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/13-s3-q12

핵심 아이디어

각 X_k는 위치 k에서 A들의 새로운 블록이 시작될 때에만 정확히 1이 되므로, S는 줄에 있는 A들의 블록(연, runs)의 개수를 센다. 이 변수들은 지시변수(indicator variable)이므로, 모든 기댓값(expectation)은 단지 몇몇 지정된 위치에 지정된 문자가 놓일 확률일 뿐이며, 이 확률은 관여하는 위치가 어디인지가 아니라 몇 개인지에만 의존한다. 분산(variance)을 구하기 위해서는 S^2을 전개하고 곱 X_iX_j를 유형별로 분류한다. 첨자가 같은 경우, 첨자가 이웃한 경우(이 경우는 불가능함이 밝혀진다), 그리고 나머지 경우이다.

풀이

전체에 걸쳐 $\binom{n}{a}$가지 배열은 모두 같은 확률로 나타난다. 이로부터 유용한 사실을 얻는다. 임의의 m개의 **서로 다른** 위치에 대해, 이 위치들이 p개의 A와 q개의 B($p+q=m$)로 이루어진 지정된 패턴을 가질 확률은 위치를 하나씩 채워 나가며 구할 수 있다. 예를 들어

$$\Pr(\text{위치} r, s\text{에 각각} B, A) = \frac{b}{n} \cdot \frac{a}{n-1},$$

이고, 이는 서로 다른 위치 r와 s가 무엇이든 성립한다. 각 X_k는 0과 1의 값만 가지므로 $\mathrm{E}(X_k) = \Pr(X_k = 1)$이고, 마찬가지로 $\mathrm{E}(X_iX_j) = \Pr(X_i = 1\text{이고}X_j = 1)$이다.

소문항 (i)

$i = 1$일 때는 첫 문자가 A이어야 하므로

$$\mathrm{E}(X_1) = \frac{a}{n}.$$

$2 \le i \le n$일 때는 위치 $i-1$에 B가, 위치 i에 A가 있어야 하므로

$$\mathrm{E}(X_i) = \frac{b}{n} \cdot \frac{a}{n-1} = \frac{ab}{n(n-1)}.$$

기댓값의 선형성(linearity of expectation)에 의해, $i \ge 2$인 $n-1$개의 첨자에 대해 더하면

$$\mathrm{E}(S) = \frac{a}{n} + (n-1) \cdot \frac{ab}{n(n-1)} = \frac{a}{n} + \frac{ab}{n} = \frac{a(b+1)}{n},$$

이므로 요구된 결과를 얻는다.

소문항 (ii)

(a) $j \ge 3$이라 하자. 그러면 $X_1 X_j = 1$인 것은 위치 1에 A, 위치 $j-1$에 B, 위치 j에 A가 있을 때와 정확히 같다. $j-1 \ge 2$이므로 이들은 서로 다른 세 위치이고, 두 개의 A와 한 개의 B를 담는다. 차례로 채우면

$$\mathrm{E}(X_1 X_j) = \frac{a}{n} \cdot \frac{b}{n-1} \cdot \frac{a-1}{n-2} = \frac{a(a-1)b}{n(n-1)(n-2)}.$$

(b) 이제 $2 \le i$이고 $j \ge i+2$라 하자. 그러면 $X_i X_j = 1$인 것은 위치 $i-1$, i, $j-1$, j에 각각 B, A, B, A가 있을 때와 정확히 같다. $j-1 \ge i+1 > i$이므로 이 네 위치는 서로 다르고, 두 개의 A와 두 개의 B를 담으므로

$$\mathrm{E}(X_i X_j) = \frac{b}{n} \cdot \frac{a}{n-1} \cdot \frac{b-1}{n-2} \cdot \frac{a-1}{n-3} = \frac{a(a-1)b(b-1)}{n(n-1)(n-2)(n-3)}.$$

(여기서 $a, b \ge 2$이므로 $n \ge 4$이고, 따라서 분모가 0이 되는 일은 없다.) 이 값은 이중합의 모든 쌍에 대해 같으므로 쌍의 개수만 세면 된다. 주어진 i에 대해 j는 $i+2$부터 n까지, 즉 $n-i-1$개의 값을 가진다. i가 2부터 $n-2$까지 변하면 이 값은 $n-3$, $n-4$, …, 1이 된다. 따라서 쌍의 개수는

$$1 + 2 + \cdots + (n-3) = \tfrac{1}{2}(n-3)(n-2),$$

이고

$$\sum_{i=2}^{n-2} \left(\sum_{j=i+2}^{n} \mathrm{E}(X_i X_j) \right) = \frac{(n-2)(n-3)}{2} \cdot \frac{a(a-1)b(b-1)}{n(n-1)(n-2)(n-3)} = \frac{a(a-1)b(b-1)}{2n(n-1)}.$$

(c) $\mathrm{Var}(S) = \mathrm{E}(S^2) - (\mathrm{E}(S))^2$을 이용하고 다음과 같이 전개한다.

$$S^2 = \sum_{i=1}^{n} X_i^2 + 2 \sum_{i<j} X_i X_j.$$

X_i는 0 또는 1이므로 $X_i^2 = X_i$이고, 따라서 첫 번째 합의 기댓값은 $\mathrm{E}(S)$이다.

? 생각해 보기 (b)에서 빠진, $j = i+1$인 쌍들은 어떻게 되는가?

이 쌍들은 아무것도 기여하지 않는다. 항상 $X_iX_{i+1} = 0$이기 때문이다. $i = 1$일 때, $X_1 = 1$이려면 위치 1에 A가 있어야 하지만 $X_2 = 1$이려면 그 자리에 B가 있어야 한다. $i \geq 2$일 때, $X_i = 1$이려면 위치 i에 A가 있어야 하지만 $X_{i+1} = 1$이려면 그 자리에 B가 있어야 한다. (A들의 두 블록이 이웃한 위치에서 시작할 수는 없다.)

따라서 의미 있는 쌍 $i < j$는 $j \geq i+2$인 것들이다. 그중 $i = 1$인 것은 $j = 3, \ldots, n$으로 $n-2$개의 쌍이며, 각각은 (a)에서 구한 값을 가진다. $i \geq 2$인 것들은 정확히 (b)의 이중합이다. 그러므로

$$\begin{aligned}\sum_{i<j} \mathrm{E}(X_iX_j) &= (n-2)\cdot\frac{a(a-1)b}{n(n-1)(n-2)} + \frac{a(a-1)b(b-1)}{2n(n-1)} \\ &= \frac{a(a-1)b}{n(n-1)}\left(1+\frac{b-1}{2}\right) = \frac{a(a-1)b(b+1)}{2n(n-1)},\end{aligned}$$

이고, 따라서

$$\mathrm{E}(S^2) = \frac{a(b+1)}{n} + \frac{a(a-1)b(b+1)}{n(n-1)}.$$

$(\mathrm{E}(S))^2 = a^2(b+1)^2/n^2$을 빼고 공통인수 $\dfrac{a(b+1)}{n^2(n-1)}$로 묶으면

$$\mathrm{Var}(S) = \frac{a(b+1)}{n^2(n-1)}\Big[n(n-1) + n(a-1)b - a(b+1)(n-1)\Big].$$

괄호를 전개하면 nab와 $-nab$가 소거되어 다음이 남는다.

$$\begin{aligned}n^2 - n - nb - na + ab + a &= n(n-a-b) - n + ab + a \\ &= -n + ab + a \qquad \text{왜냐하면} n = a+b \\ &= ab - b = b(a-1).\end{aligned}$$

따라서

$$\mathrm{Var}(S) = \frac{a(b+1)}{n^2(n-1)}\cdot b(a-1) = \frac{a(a-1)b(b+1)}{n^2(n-1)},$$

이므로 요구된 결과를 얻는다.

검산

$a = 3, b = 2$로 두면 $n = 5$이고 배열은 10가지이다. A들의 블록을 세어 보면, 배열 $AAABB$, $BAAAB$, $BBAAA$는 $S = 1$을, $ABABA$는 $S = 3$을, 나머지 여섯 개는 $S = 2$를 준다. 따라서 $\mathrm{E}(S) = \frac{18}{10} = \frac{9}{5}$이고 $\mathrm{E}(S^2) = \frac{36}{10}$이므로 $\mathrm{Var}(S) = \frac{18}{5} - \frac{81}{25} = \frac{9}{25}$이다. 공식은 $\frac{3\cdot3}{5} = \frac{9}{5}$와 $\frac{3\cdot2\cdot2\cdot3}{25\cdot4} = \frac{9}{25}$를 준다. 컴퓨터로 모든 경우를 나열하여, 모든 $2 \leq a, b \leq 6$에 대해 문제의 모든 공식이 정확히 성립함을 확인하였다.

이 공식들은 주어진 범위 밖에서도 타당하게 작동한다. $a = 2, b = 1$일 때 배열 AAB, ABA, BAA는 $S = 1, 2, 1$을 주므로 $\mathrm{E}(S) = \frac{4}{3}$이고 $\mathrm{Var}(S) = 2 - \frac{16}{9} = \frac{2}{9}$이며, 이는 $\frac{2\cdot 2}{3}$ 및 $\frac{2\cdot 1\cdot 1\cdot 2}{9\cdot 2}$와 일치한다. 또 $a = 1$이면 분산 공식은 0을 주는데, 이때는 블록이 항상 정확히 하나이므로 마땅히 그래야 한다.

흔한 실수

- X_1을 다른 변수들과 똑같이 취급하는 것. 위치 0은 없으므로 X_1은 문자 하나만 필요하고, $\mathrm{E}(X_1) = a/n$이지 $ab/(n(n-1))$이 아니다.
- X_i와 X_{i+1}이 동시에 1일 수 없다는 것을 잊거나, $\mathrm{E}(X_i X_{i+1})$을 0이 아닌 식으로 계산하는 것.
- (b)에서 쌍의 개수를 잘못 세는 것. 쌍은 $\frac{1}{2}(n-2)(n-3)$개이지 $\frac{1}{2}n(n-1)$개가 아니다.

STEP 너머

어떤 개수를 지시변수들의 합으로 쓴 다음, 평균은 선형성으로, 분산은 쌍별 곱으로 구하는 방법은 대학 확률론에서 가장 유용한 기법 중 하나이다. 연, 짝짓기(matching), 무작위 순열(random permutation)의 고정점(fixed point), 무작위 그래프(random graph)를 분석하는 방법이 바로 이것이다. 여기서 S는 본질적으로 수열의 비무작위성을 탐지하는 데 쓰이는 왈드–울포위츠 연 검정(Wald–Wolfowitz runs test)의 통계량이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에는 지시변수 방법의 예가 많이 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.