풀이

STEP III 2004, 문제 12

확률과 통계 • 지시변수와 교란순열

> **STEP III 2004, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/04-s3-q12

핵심 아이디어

개수를 지시확률변수(indicator variable)의 합으로 나타내는 것은 확률론에서 가장 유용한 기법 중 하나이다. 지시확률변수의 기댓값(expectation)은 그냥 확률이므로 $\mathrm{E}(N)$은 거의 저절로 구해진다. 분산(variance)을 구하려면 쌍들 사이의 공분산(covariance)이 필요한데, 셔츠가 독립적으로 배정되지 않기 때문에 이 값들은 작지만 0은 아니다. N의 평균과 분산이 모두 1임을 알고 나면, 모수가 1인 푸아송 분포(Poisson distribution)가 시도해 볼 만한 자연스러운 모형이 된다.

풀이

지시확률변수 하나, C_1

셔츠를 배정하는 $m!$가지 방법은 모두 같은 확률로 일어나므로, 셔츠 1은 m명의 선수 중 누구에게나 같은 확률로 돌아간다. 따라서 $\mathrm{P}(C_1 = 1) = \frac{1}{m}$이고, C_1은 0과 1의 값만 가지므로

$$\mathrm{E}(C_1) = 1 \cdot \mathrm{P}(C_1 = 1) = \frac{1}{m}$$

이 되어 요구된 결과를 얻는다. 또한 $C_1^2 = C_1$이므로($0^2 = 0$이고 $1^2 = 1$이므로) $\mathrm{E}(C_1^2) = \frac{1}{m}$이고, 따라서

$$\mathrm{Var}(C_1) = \frac{1}{m} - \frac{1}{m^2} = \frac{m-1}{m^2}$$

이다.

C_1과 C_2의 공분산

여기서는 $m \geq 2$가 필요하다. 곱 C_1C_2는 선수 1과 선수 2가 둘 다 자기 셔츠를 받을 때 정확히 1이고, 그렇지 않으면 0이다. 선수 1이 셔츠 1을 받았다고 하면, 남은 $m-1$장의 셔츠는 나머지 $m-1$명의 선수에게 무작위로 나누어지므로 선수 2가 셔츠 2를 받을 확률은 $\frac{1}{m-1}$이다. 그러므로

$$\mathrm{E}(C_1C_2) = \mathrm{P}(C_1 = 1,\, C_2 = 1) = \frac{1}{m} \cdot \frac{1}{m-1}$$

이고

$$\begin{aligned}\mathrm{Cov}(C_1, C_2) &= \mathrm{E}(C_1C_2) - \mathrm{E}(C_1)\,\mathrm{E}(C_2) = \frac{1}{m(m-1)} - \frac{1}{m^2} \\ &= \frac{m-(m-1)}{m^2(m-1)} = \frac{1}{m^2(m-1)}\end{aligned}$$

이다. 이 값은 양수이다. 선수 1이 맞는 셔츠를 받았다는 사실을 알면 선수 2에게 돌아갈 수 있는 셔츠 더미에서 틀린 셔츠 하나가 빠지므로, 선수 2의 확률이 조금 올라가기 때문이다.

> **생각해 보기** 공분산이 왜 0이 아닐까? 선수들의 셔츠는 따로따로 정해지는 것이 아닌가?

그렇지 않다. 셔츠는 선수마다 한 장씩 배정되므로, 선수 1이 무엇을 받느냐가 다른 모든 선수에게 남는 셔츠에 영향을 준다. C_i들은 **종속**(*dependent*)이며, 바로 그 때문에 $\mathrm{Var}(N)$에 공분산 항들을 포함해야 한다.

N의 평균과 분산

대칭성에 의해 모든 C_i는 C_1과 같은 분포를 가지고, $i \neq j$인 모든 쌍 C_i, C_j는 C_1, C_2와 같은 결합분포를 가진다(선수와 셔츠의 번호를 함께 바꾸어 붙여도 무작위 배정은 달라지지 않는다). 독립성이 전혀 필요 없는 기댓값의 선형성(linearity of expectation)에 의해

$$\mathrm{E}(N) = \sum_{i=1}^{m} \mathrm{E}(C_i) = m \cdot \frac{1}{m} = 1$$

이다. 합의 분산을 구하려면 모든 분산과, $i \neq j$인 순서쌍 전체에 대한 모든 공분산을 더한다. 이러한 순서쌍은 $m(m-1)$개이다.

$$\begin{aligned}\mathrm{Var}(N) &= \sum_{i=1}^{m} \mathrm{Var}(C_i) + \sum_{i \neq j} \mathrm{Cov}(C_i, C_j) \\ &= m \cdot \frac{m-1}{m^2} + m(m-1) \cdot \frac{1}{m^2(m-1)} \\ &= \frac{m-1}{m} + \frac{1}{m} = 1.\end{aligned}$$

따라서 요구된 대로 $\mathrm{E}(N) = \mathrm{Var}(N) = 1$이다. ($m = 1$이면 항상 $N = 1$이므로 $\mathrm{Var}(N) = 0$이 된다. 공분산을 구할 때 이미 필요했듯이, 문제는 암묵적으로 $m \geq 2$를 가정하고 있다.)

어떤 근사를 쓸까?

정규 근사(Normal approximation)는 m의 값이 무엇이든 좋지 않다. N은 정수 값 $0, 1, 2, \ldots$만 가지며, 평균이 1이고 표준편차(standard deviation)가 1이므로 확률의 거의 전부가 $0, 1, 2, 3$의 서너 개 값에 몰려 있다. 분포는 평균보다 표준편차 하나만큼 아래인 0에서 잘려 있으므로 오른쪽으로 크게 치우쳐(skewed) 있다. 반면 $\mathrm{N}(1, 1)$ 분포는 대칭이고, 연속성 수정(continuity correction)을 하더라도 N이 결코 가질 수 없는 $-\frac{1}{2}$ 아래에 무시할 수 없는 확률(약 7%)을 준다. 결정적으로, m을 늘려도 도움이 되지 않는다. 평균과 분산은 모든 m에 대해 1로 유지되므로, 분포의 모양이 퍼져서 종 모양이 되는 일은 결코 없다.

반면 푸아송 근사는 그럴듯하다. N은 각각 작은 확률 $\frac{1}{m}$을 가지는 많은 수 m개의 사건 가운데 "성공"의 개수를 세며, 이 사건들은 약하게만 종속되어 있다(공분산 $\frac{1}{m^2(m-1)}$은 매우 작다). 바로 이런 상황에서 개수가 근사적으로 푸아송 분포를 따른다. 게다가 푸아송 확률변수는 평균과 분산이 같은데, N이 정확히 이 성질을 가지므로 $\mathrm{Po}(1)$이 자연스러운 선택이다.

$m = 4$인 경우

같은 확률로 일어나는 배정은 $4! = 24$가지이다. 배정을 선수 $1, 2, 3, 4$가 입은 셔츠의 목록으로 나타내면, 아무도 자기 셔츠를 받지 않는 배정은

$$2143, \quad 2341, \quad 2413, \quad 3142, \quad 3412,$$
$$3421, \quad 4123, \quad 4312, \quad 4321.$$

이다. (체계적으로 나열하려면, 선수 1이 셔츠 2, 3, 4 중 하나를 받고, 각 경우마다 나머지를 마무리하는 방법이 정확히 세 가지이므로 $3 \times 3 = 9$이다.) 따라서

$$\mathrm{P}(N = 0) = \frac{9}{24} = \frac{3}{8} = 0.375$$

이다.

모수가 1인 푸아송 근사는 $\mathrm{P}(N = 0) \approx \mathrm{e}^{-1}$을 준다. $\mathrm{e} \approx 2.72$를 쓰면

$$\mathrm{e}^{-1} \approx \frac{100}{272} = \frac{25}{68}$$

이다. 참값을 기준으로 잰 상대오차(relative error)는

$$\frac{\frac{3}{8} - \frac{25}{68}}{\frac{3}{8}} = 1 - \frac{25}{68} \cdot \frac{8}{3} = 1 - \frac{200}{204} = \frac{1}{51} \approx 0.0196$$

이고, 이는 주장한 대로 약 2%이다. (대신 근사값을 기준으로 오차를 재면 $\frac{1}{50}$, 즉 정확히 2%가 된다. 어느 쪽이든 이렇게 작은 m에 대해 놀라울 만큼 잘 맞는다.)

검산

- $m = 2$이면 가능한 경우는 둘 다 맞는 경우($N = 2$)와 둘 다 틀린 경우($N = 0$)뿐이고, 각각의 확률은 $\frac{1}{2}$이다. 평균은 1, 분산은 1이다. 맞다.
- $m = 2, \ldots, 7$에 대해 모든 순열을 컴퓨터로 나열하여 $\mathrm{E}(C_1) = \frac{1}{m}$, $\mathrm{Var}(C_1) = \frac{m-1}{m^2}$, $\mathrm{Cov}(C_1, C_2) = \frac{1}{m^2(m-1)}$, $\mathrm{E}(N) = \mathrm{Var}(N) = 1$이 정확히 성립함을 확인했고, 4개 대상의 교란순열(derangement)이 9개임을 찾았다. 따라서 $\mathrm{P}(N = 0) = 0.375$이고, 이를 $1/2.72 \approx 0.3676$과 비교할 수 있다.
- $N = m - 1$은 불가능하다는 점에 주목하자. 한 명을 뺀 모든 선수가 맞는 셔츠를 받았다면, 마지막 선수도 맞는 셔츠를 받는다. 이것도 N이 정규분포와 매우 거리가 멀다는 또 하나의 신호이다.

흔한 실수

- C_i들을 독립으로 취급하는 것. 그러면 1 대신 $\mathrm{Var}(N) = \frac{m-1}{m}$을 얻는다.
- 공분산 항을 $m(m-1)$개가 아니라 $\binom{m}{2}$개만 세는 것. 즉 $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$에서 인수 2를 빠뜨리는 것이다.
- 교란순열을 부주의하게 나열하는 것. 체계적으로 세면(3×3) 하나를 빠뜨리는 일을 피할 수 있다.

STEP 너머

이것은 고전적인 "맞추기(matching)" 문제, 곧 교란순열 문제이다. 포함배제 원리(inclusion–exclusion)를 이용하면 $\mathrm{P}(N = 0) = \sum_{k=0}^{m} \frac{(-1)^k}{k!}$임을 보일 수 있고, 이 값은 e^{-1}로 매우 빠르게 수렴한다. 실제로 $m \to \infty$일 때 N은 $\mathrm{Po}(1)$로 분포수렴(convergence in distribution)한다. 드물고 약하게 종속된 많은 사건의 개수는 거의 푸아송 분포를 따른다는 더 넓은 원리는 첸–스타인 방법(Chen–Stein method)으로 엄밀하게 다듬어진다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 지시확률변수와 맞추기 문제를 반복해서 등장하는 예로 사용한다.

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.