풀이

STEP III 1988, 문제 16

확률과 통계 • 초기하분포

STEP III 1988, 16번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/88-s3-q16

핵심 아이디어

비복원추출(drawing without replacement)에서는 각 시행이 서로 종속(dependent)이므로, 한 번씩 차례로 따져 나가면 계산이 복잡해진다. 이 어려움은 두 가지 아이디어로 피해 갈 수 있다. 첫째, 순서를 고려하지 않은 선택의 수를 세어 빨간 공의 개수의 분포(distribution)를 구한다. 둘째, 대칭성(symmetry)을 이용한다. 항아리 속의 공을 무작위 순서로 모두 꺼낸다고 상상하면, 각 공이 주어진 어느 위치에 놓일 가능성은 모두 같다. 그러면 X를 지시변수(indicator variable) X_i들의 합으로 나타내어, 분포를 전혀 구하지 않고도 평균(mean)과 분산(variance)을 얻을 수 있다. 필요한 것은 두 변수 사이의 공분산(covariance)뿐이다.

풀이

빨간 공이 정확히 k개일 확률

공들을 서로 구별할 수 있다고 생각하자. 꺼낸 n개의 공은 크기가 n인 집합을 이루고, 이러한 집합은 모두 같은 확률로 나타나므로 같은 확률의 결과가 $\binom{M}{n}$가지 있다. 빨간 공이 정확히 k개인 집합은 m개의 빨간 공 중 k개를, $M-m$개의 초록 공 중 $n-k$개를 골라 만든다. 따라서

$$\mathrm{P}(\text{정확히}k\text{개가 빨간 공}) = \frac{\binom{m}{k}\binom{M-m}{n-k}}{\binom{M}{n}},$$

이다. 여기서 $b < 0$ 또는 $b > a$이면 $\binom{a}{b} = 0$이라는 일반적인 약속을 따른다. (이 약속으로 $n-k > M-m$과 같은 불가능한 경우도 함께 다룰 수 있다.) 이것이 **초기하분포**(*hypergeometric distribution*)이다.

소문항 (i)

항아리가 빌 때까지 계속 공을 꺼낸다고 상상하자. (구별 가능한) M개의 공이 나오는 순서는 균등한 무작위 배열이다. 즉 $M!$가지 순서가 모두 같은 확률로 나타난다. n번 꺼낸 뒤 멈추더라도 처음 n개의 위치는 달라지지 않으므로, 전체 배열을 가지고 생각해도 된다.

위치 i를 하나 고정하자. 특정한 공이 위치 i에 오는 배열의 수는 $(M-1)!$로, 어느 공이든 같다. 따라서 위치 i에 있는 공은 M개의 공 중 어느 것일 가능성이나 모두 같고,

$$\mathrm{P}(X_i = 1) = \frac{m\,(M-1)!}{M!} = \frac{m}{M},$$

이므로 원하는 결과가 성립한다.

> **생각해 보기** 두 번째 공의 색은 첫 번째 공이 무엇이었는지에 따라 달라지지 않는가?

첫 번째 공이 주어졌을 때는 그렇다. 그러나 우리가 구하려는 것은 **조건 없는**(*unconditional*) 확률이다. $i = 2$일 때 전확률 법칙(law of total probability)을 쓰면

$$\mathrm{P}(X_2 = 1) = \frac{m}{M}\cdot\frac{m-1}{M-1} + \frac{M-m}{M}\cdot\frac{m}{M-1} = \frac{m(M-1)}{M(M-1)} = \frac{m}{M}.$$

이다. 대칭성 논증은 이와 같은 계산을 모든 i에 대해 한꺼번에 해 준다.

소문항 (ii)

서로 다른 위치 $i \neq j$를 고정하자 (따라서 $M \geq 2$이다). 두 위치 모두에 빨간 공이 오는 배열의 수를 세자. 위치 i의 공을 고르는 방법이 m가지, 그다음 위치 j의 공을 고르는 방법이 $m-1$가지이고, 나머지를 배열하는 방법이 $(M-2)!$가지이다. 따라서

$$\mathrm{P}(X_i = 1 \text{ 그리고} X_j = 1) = \frac{m(m-1)(M-2)!}{M!} = \frac{m(m-1)}{M(M-1)},$$

이므로 원하는 결과가 성립한다. 다시 말해, 위치 (i, j)에 있는 공의 순서쌍은 서로 다른 두 공으로 이루어진 $M(M-1)$개의 순서쌍 중 어느 것일 가능성이나 모두 같다.

X의 평균과 분산

X는 처음 꺼낸 n개의 공 중 빨간 공의 개수이다. 독립성이 전혀 필요 없는 기댓값의 선형성(linearity of expectation)과 $\mathrm{E}(X_i) = \mathrm{P}(X_i = 1)$을 이용하면

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{nm}{M}.$$

이다.

분산은 $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$을 이용하여 구한다. 합의 제곱을 전개하고, X_i는 0 또는 1이므로 $X_i^2 = X_i$임을 이용하면

$$X^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j = X + \sum_{i \neq j} X_i X_j,$$

이다. 여기서 두 번째 합은 $i \neq j$인 $n(n-1)$개의 순서쌍 (i, j) 전체에 대한 합이다. 이제 $X_i X_j = 1$인 것은 두 공이 모두 빨간 공일 때뿐이므로, 소문항 (ii)에 의해 $\mathrm{E}(X_i X_j) = \dfrac{m(m-1)}{M(M-1)}$이다. 따라서 $M \geq 2$일 때

$$\begin{aligned}\mathrm{Var}(X) &= \frac{nm}{M} + \frac{n(n-1)m(m-1)}{M(M-1)} - \frac{n^2 m^2}{M^2} \\ &= \frac{nm}{M^2(M-1)}\Big[M(M-1) + (n-1)(m-1)M - nm(M-1)\Big].\end{aligned}$$

이다. 대괄호 안을 전개하면

$$\begin{aligned}&M^2 - M + nmM - nM - mM + M - nmM + nm \\ &\quad = M^2 - nM - mM + nm = (M-m)(M-n).\end{aligned}$$

이다. 따라서

$$\mathrm{Var}(X) = \frac{nm(M-m)(M-n)}{M^2(M-1)}.$$

이다. ($M = 1$이면 $n \leq 1$이므로 순서쌍이 없고 X는 상수이므로 $\mathrm{Var}(X) = 0$이다.)

$p = m/M$으로 놓으면 많은 것이 드러난다. 이때

$$\mathrm{E}(X) = np, \qquad \mathrm{Var}(X) = np(1-p) \cdot \frac{M-n}{M-1}.$$

이다. 이것은 이항분포(binomial distribution)의 분산에 보정 인수를 곱한 것이며, 이 인수는 $n \geq 2$일 때 1보다 작다.

검산

- $n = M$: 모든 공을 꺼내므로 반드시 $X = m$이다. 공식에서도 평균은 m이고, 인수 $M - n$ 때문에 분산은 0이다.
- $n = 1$: X는 $p = m/M$인 하나의 베르누이(Bernoulli) 확률변수이고, 분산은 $p(1-p)$가 된다.
- $M = 6, m = 2, n = 3$: 공식에 따르면 $\mathrm{Var}(X) = \frac{3 \cdot 2 \cdot 4 \cdot 3}{36 \cdot 5} = \frac{2}{5}$이다. 6!가지 순서를 모두 나열하여 정확히 계산한 값과 일치하며, $M \leq 6$인 모든 경우에도 일치한다.

흔한 실수

- X_i들을 독립으로 취급하여 이항분포의 분산 $np(1-p)$를 얻는 것. 이들은 음의 상관관계(negatively correlated)에 있다. 앞에서 빨간 공이 나오면 뒤에서 빨간 공이 나올 가능성이 줄어든다.
- 순서쌍을 생각할 때 이중합 $\sum_{i \neq j}$의 항이 $\binom{n}{2}$개가 아니라 $n(n-1)$개임을 잊는 것.

STEP 너머

개수를 지시변수들의 합으로 나타내는 것은 확률론에서 가장 강력한 기법 중 하나이다. 분포 자체를 적어 내기가 절망적으로 어려울 때에도 평균과 분산을 구할 수 있게 해 준다. 인수 $\frac{M-n}{M-1}$은 표본조사(survey sampling)에서 쓰이는 **유한 모집단 수정**(*finite population correction*)이다. 이 인수를 보면 1000명을 대상으로 한 여론조사가 한 나라에 대해 주는 정보가 무한 모집단에 대해 줄 정보와 거의 같은 이유를 알 수 있다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 책 전체에 걸쳐 지시변수를 활용하며 초기하분포를 자세히 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.