풀이

STEP I 2003, 문제 12

확률과 통계 • 곱의 기댓값

> **STEP I 2003, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/03-s1-q12

핵심 아이디어

기댓값(expected value)은 모든 같은 확률의 결과에 대하여 (값 $\times$ 확률)을 더한 것이므로, 두 소문항 모두 수의 순서쌍을 나열하고 그 곱을 더하는 문제로 귀결된다. 이중합(double sum) $\sum_r \sum_s rs$는 $(\sum r)^2$으로 인수분해되며, 이것이 소문항 (i)의 전부이다. 소문항 (ii)에서 달라지는 것은 $r = s$인 쌍이 더 이상 나올 수 없다는 점뿐이므로, 같은 이중합에서 대각선 항들을 빼면 된다.

풀이

첫 번째 공에 적힌 수를 X, 두 번째 공에 적힌 수를 Y라 하자.

소문항 (i)

첫 번째 공을 다시 넣으므로, 두 번째 공은 다시 가득 찬 주머니에서 뽑히며 첫 번째 시행과 무관하다. 따라서 임의의 $r, s \in \{1, \ldots, n\}$에 대하여

$$\mathrm{P}(X = r,\ Y = s) = \mathrm{P}(X = r)\,\mathrm{P}(Y = s) = \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n^2}.$$

이러한 순서쌍 (r, s)는 n^2개이고 각각의 곱은 rs이므로, 기댓값의 정의에 의하여

$$\mathrm{E}(XY) = \sum_{r=1}^{n} \sum_{s=1}^{n} rs\,\mathrm{P}(X = r,\ Y = s) = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} rs,$$

이며, 이것이 바로 설명하라고 요구된 식이다.

이를 간단히 하기 위해, r는 s에 의존하지 않으므로 안쪽 합 밖으로 꺼낼 수 있음에 주목한다.

$$\sum_{r=1}^{n} \sum_{s=1}^{n} rs = \sum_{r=1}^{n} r\left(\sum_{s=1}^{n} s\right) = \left(\sum_{r=1}^{n} r\right)^2 = \tfrac{1}{4} n^2 (n+1)^2.$$

그러므로

$$\mathrm{E}(XY) = \frac{(n+1)^2}{4}.$$

(같은 이야기로, X와 Y는 독립(independent)이므로 $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y) = (\frac{n+1}{2})^2$이다.)

소문항 (ii)

공을 두 개 뽑으려면 $n \geq 2$이어야 한다. 이번에는 첫 번째 공을 다시 넣지 않으므로, 가능한 결과는 $r \neq s$인 순서쌍 (r, s)이다. 이러한 쌍에 대하여

$$\mathrm{P}(X = r,\ Y = s) = \mathrm{P}(X = r)\,\mathrm{P}(Y = s \mid X = r) = \frac{1}{n} \cdot \frac{1}{n-1},$$

이다. 공 r를 꺼낸 뒤에는 두 번째 공이 남은 $n-1$개 중 어느 것이든 같은 확률로 뽑히기 때문이다. 이러한 쌍은 $n(n-1)$개이고 모두 같은 확률이며, $r = s$인 쌍의 확률은 0이다. 따라서

$$\mathrm{E}(XY) = \frac{1}{n(n-1)} \sum_{r \neq s} rs.$$

> **생각해 보기** $r \neq s$인 쌍에 대해서만 rs를 어떻게 더할 수 있을까?

소문항 (i)에서 구한 **모든** 쌍에 대한 합에서, $r \cdot r = r^2$을 기여하는 대각선 항 $r = s$를 빼면 된다.

$$\sum_{r \neq s} rs = \sum_{r=1}^{n} \sum_{s=1}^{n} rs - \sum_{r=1}^{n} r^2 = \tfrac{1}{4}n^2(n+1)^2 - \tfrac{1}{6}n(n+1)(2n+1).$$

공통인수 $\frac{1}{12}n(n+1)$을 묶어 내면

$$\begin{aligned} \sum_{r \neq s} rs &= \tfrac{1}{12}n(n+1)[3n(n+1) - 2(2n+1)] \\ &= \tfrac{1}{12}n(n+1)(3n^2 - n - 2) \\ &= \tfrac{1}{12}n(n+1)(n-1)(3n+2). \end{aligned}$$

$n \geq 2$이므로 0이 아닌 $n(n-1)$로 나누면

$$\mathrm{E}(XY) = \frac{(n+1)(3n+2)}{12},$$

이 되어 요구된 결과를 얻는다.

검산

$n = 3$이라 하자. 복원추출(with replacement)에서는 아홉 개의 곱의 평균이 $\frac{1}{9}(1+2+3)^2 = 4$이고, $\frac{(3+1)^2}{4} = 4$이다. 비복원추출(without replacement)에서는 순서 없는 쌍

$\{1,2\},\{1,3\},\{2,3\}$의 곱이 $2,3,6$이고 평균은 $\frac{11}{3}$이며, $\frac{4\cdot 11}{12} = \frac{11}{3}$이다. $n = 2$이면 비복원추출에서 가능한 곱은 2뿐이고, 공식은 $\frac{3\cdot 8}{12} = 2$를 준다. 또한 12 이하의 모든 n에 대하여 정확한 전수 나열로 두 공식을 모두 확인하였다.

두 답을 비교하는 것도 좋은 타당성 점검이 된다.

$$\frac{(n+1)(3n+2)}{12} - \frac{(n+1)^2}{4} = -\frac{n+1}{12} < 0.$$

비복원추출일 때의 답이 더 작은데, 이는 $n \cdot n$과 같은 큰 대각선 곱들이 빠졌으므로 당연하다.

흔한 실수

- 소문항 (ii)에서 n^2으로 나누는 것. 가능한 순서쌍은 $n(n-1)$개뿐이고, 각각의 확률은 $\frac{1}{n(n-1)}$이다.
- 합은 순서 없는 쌍으로 구하고 개수는 순서쌍으로 세는 것(또는 그 반대). 순서 없는 쌍에 대해 더하면 합과 개수가 모두 절반이 되므로 비는 같지만, 둘을 섞으면 2배만큼 어긋난다.
- 소문항 (ii)에서 $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$라고 쓰는 것. 이 식은 독립성이 필요한데, 공을 다시 넣지 않으면 X와 Y는 독립이 아니다.

STEP 너머

우리가 구한 차이는 정확히 공분산(covariance)이다. 즉 $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = -\frac{n+1}{12}$이며, 첫 번째 수가 크면 두 번째 수가 클 가능성이 조금 줄어들기 때문에 음수가 된다. 이러한 공분산은 비복원추출에서 합의 분산(variance)을 구할 때 필요하며, 이로부터 초기하분포(hypergeometric distribution)의 분산과 표본조사에서 쓰이는 "유한모집단 수정(finite population correction)"이 나온다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 바로 이런 종류의 공분산과 지시변수(indicator) 방법을 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.