풀이

STEP III 2021, 문제 12

확률과 통계 • 합의 분산

> **STEP III 2021, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 OCR에서 내려받을 수 있는 2021년 STEP 시험지 묶음(zip 파일) 가운데 STEP III 시험지에 있다:
>
> www.ocr.org.uk/Images/696588-2021-past-papers.zip

이 문제는 *Beyond STEP*의 12.4 절 **둘씩은 독립이지만 한꺼번에는 독립이 아닌**에서 단계별 안내와 함께 자세히 다룬다: https://beyondstep.keithmcnulty.org/ko/ch12/pairwise-independence.html

핵심 아이디어

팀의 점수를 선수 쌍마다 하나씩 대응하는 지시 변수(indicator variable) 꼴의 확률변수들의 합으로 쓰고, 그 합의 분산(variance)을 전개한다. 그러면 모든 것은 **서로 다른** 두 쌍이 어떻게 상호작용하는지에 달려 있다. 두 쌍에 공통인 선수가 없으면 분명히 독립(independent)이므로, 실제로 다뤄야 할 것은 선수 한 명을 공유하는 쌍들이다. (i)에서는 그런 쌍들이 독립임이 밝혀지므로 교차항이 사라진다. (iii)에서는 독립이 아니며, (ii)의 항등식이 그런 쌍들의 기여를 세는 도구가 된다.

풀이

소문항 (i)

각 주사위는 공정하고 선수들은 독립적으로 주사위를 던진다. 주어진 두 선수가 같은 수를 낼 확률은 첫 번째 선수가 무엇을 내든 $\frac{1}{6}$이다. 따라서

$$\mathrm{P}(X_{12}=1)=\mathrm{P}(X_{23}=1)=\tfrac{1}{6}.$$

$X_{12}=1$과 $X_{23}=1$이 모두 일어나는 것은 정확히 선수 1, 2, 3이 모두 같은 수를 낼 때이므로

$$\mathrm{P}(X_{12}=1,\ X_{23}=1)=\tfrac{1}{6}\cdot\tfrac{1}{6}=\tfrac{1}{36}=\mathrm{P}(X_{12}=1)\,\mathrm{P}(X_{23}=1).$$

이 값을 주변 확률(marginal probability)에서 빼면 나머지 세 칸을 얻는다. 예를 들어 $\mathrm{P}(X_{12}=1,X_{23}=0)=\frac{1}{6}-\frac{1}{36}=\frac{1}{6}\cdot\frac{5}{6}$이고, 마찬가지로 $\mathrm{P}(X_{12}=0,X_{23}=1)=\frac{5}{6}\cdot\frac{1}{6}$,

$\mathrm{P}(X_{12}=0, X_{23}=0) = 1-\frac{1}{6}-\frac{1}{6}+\frac{1}{36} = \frac{5}{6}\cdot\frac{5}{6}$이다. 따라서 결합 분포(joint distribution)는 주변 분포들의 곱이고, X_{12}와 X_{23}은 독립이다.

> ? **생각해 보기** $X_{12}=1$임을 알아도 왜 X_{23}에 대해 아무것도 알 수 없는가?

선수 1과 2가 같은 수를 냈다는 조건 아래에서도 선수 2의 수는 여전히 $\{1,\dots,6\}$ 위에서 균등하게 분포하고, 선수 3이 그 수와 같을 확률은 이전과 똑같이 $\frac{1}{6}$이다. 모든 눈이 똑같이 나올 수 있으므로 공유된 선수는 아무 정보도 전달하지 않는다.

$T=\sum_{i<j} X_{ij}$를 총점이라 하자. 이는 $N=\binom{n}{2}$개 항의 합이다. 각 X_{ij}는 모수가 $\frac{1}{6}$인 베르누이(Bernoulli) 확률변수이므로 $\mathrm{E}(X_{ij})=\frac{1}{6}$이고 $\mathrm{Var}(X_{ij})=\frac{1}{6}\cdot\frac{5}{6}=\frac{5}{36}$이다. 그러므로

$$\mathrm{E}(T)=\frac{N}{6}=\frac{n(n-1)}{12}.$$

분산을 구하기 위해 서로 다른 두 쌍을 아무거나 택하자. 공통인 선수가 없으면 두 쌍은 서로소인 주사위 집합에 의존하므로 독립이다. 선수 한 명을 공유하면, 선수 번호를 다시 붙인 뒤 위의 논증에 의해 독립이다. 따라서 X_{ij}들은 쌍별 독립이고, 모든 공분산(covariance)이 0이며, 합의 분산은 분산들의 합이다.

$$\mathrm{Var}(T)=N\cdot\frac{5}{36}=\frac{5n(n-1)}{72}.$$

소문항 (ii)

$S=Y_1+\cdots+Y_m$이라 하자. 선형성(linearity)에 의해 $\mathrm{E}(S)=0$이므로 $\mathrm{Var}(S)=\mathrm{E}(S^2)-0^2=\mathrm{E}(S^2)$이다. 제곱을 전개하면 각 Y_i^2은 한 번, $i<j$인 각 곱 Y_iY_j는 두 번 나타난다.

$$S^2=\sum_{i=1}^{m} Y_i^2+2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} Y_iY_j.$$

기댓값(expectation)을 취하고 선형성을 이용하면 구하는 공식을 얻는다.

소문항 (iii)

두 선수에 대해 $\mathrm{P}(\text{같은 짝수})=\frac{3}{36}=\frac{1}{12}$이고, 마찬가지로 $\mathrm{P}(\text{같은 홀수})=\frac{1}{12}$이다. 따라서

$$\mathrm{P}(Z_{ij}=1)=\mathrm{P}(Z_{ij}=-1)=\tfrac{1}{12},\qquad \mathrm{P}(Z_{ij}=0)=\tfrac{5}{6}.$$

독립이 아님. $Z_{12}=1$이고 $Z_{23}=-1$이라면 선수 2의 수는 짝수이면서 동시에 홀수여야 한다. 따라서

$$\mathrm{P}(Z_{12}=1,\ Z_{23}=-1)=0\neq\tfrac{1}{12}\cdot\tfrac{1}{12}=\mathrm{P}(Z_{12}=1)\,\mathrm{P}(Z_{23}=-1),$$

이고, Z_{12}와 Z_{23}은 독립이 아니다.

평균. 대칭성에 의해 $\mathrm{E}(Z_{ij}) = \frac{1}{12} - \frac{1}{12} = 0$이므로 총점 $U = \sum_{i<j} Z_{ij}$의 평균은 0이다.

분산. 모든 Z_{ij}의 평균이 0이므로 $N = \binom{n}{2}$개의 변수 Z_{ij}에 (ii)를 쓸 수 있다. 두 종류의 기댓값이 필요하다.

- 제곱: Z_{ij}^2은 정확히 선수 i와 j가 같은 수를 낼 때 1이므로 $\mathrm{E}(Z_{ij}^2) = \frac{1}{6}$이다.
- 서로 다른 두 쌍의 곱. 두 쌍에 공통인 선수가 없으면 두 변수는 평균이 0이고 독립이므로 $\mathrm{E}(Z_{ij}Z_{kl}) = \mathrm{E}(Z_{ij})\,\mathrm{E}(Z_{kl}) = 0$이다. 선수를 공유하면, 예를 들어 $Z_{12}Z_{23}$의 경우 세 선수가 모두 같은 수를 낼 때에만 곱이 0이 아니다. 이때 두 인수는 모두 $+1$(짝수)이거나 모두 -1(홀수)이므로 곱은 1이다. 그러므로 $\mathrm{E}(Z_{12}Z_{23}) = \frac{1}{36}$이다.

? **생각해 보기** 정확히 한 명의 선수를 공유하는 쌍들의 (순서 없는) 쌍은 몇 개인가?

공유되는 선수를 n가지 방법으로 고르고, 그 선수의 두 짝을 나머지 $n-1$명 중에서 $\binom{n-1}{2}$가지 방법으로 고른다. 그러면 이런 쌍들의 쌍이 $n\binom{n-1}{2} = \frac{1}{2}n(n-1)(n-2)$개이고 각각 한 번씩 세어지는데, 이것이 바로 (ii)의 이중합에 필요한 것이다.

이를 종합하면

$$\begin{aligned}\mathrm{Var}(U) &= \binom{n}{2}\cdot\frac{1}{6} + 2\cdot\frac{n(n-1)(n-2)}{2}\cdot\frac{1}{36}\\ &= \frac{n(n-1)}{36}\big(3+(n-2)\big)\\ &= \frac{n(n-1)(n+1)}{36} = \frac{1}{36}n(n^2-1),\end{aligned}$$

이므로 증명이 끝난다.

검산

- 예시로 주어진 결과 $3,3,3,3,6,6,2$에 대해 쌍의 개수는 $6+1=7$, 부호를 붙여 센 값은 $-6+1=-5$로 문제와 일치한다.
- $n=2$이면 쌍이 하나뿐이므로 $\mathrm{Var}(U) = \mathrm{E}(Z_{12}^2) = \frac{1}{6}$이고, 공식도 $\frac{2\cdot3}{36} = \frac{1}{6}$을 준다.
- $n=3$일 때, 세 주사위가 모두 같으면 $U = \pm3$이다(각각 확률 $\frac{1}{72}$). 그렇지 않으면 정확히 한 쌍만 같고 $U = \pm1$이며, 그 확률의 합은 $\frac{15}{36}$이다. 따라서 $\mathrm{E}(U^2) = 9\cdot\frac{1}{36} + \frac{15}{36} = \frac{2}{3}$으로 $\frac{3\cdot8}{36}$과 일치한다. $n \le 6$에 대해 모든 결과를 컴퓨터로 정확히 열거해 보면 두 평균과 두 분산이 모두 확인된다. 예를 들어 $n=6$이면 (i)에서 $\frac{25}{12}$, (iii)에서 $\frac{35}{6}$이다.

흔한 실수

- 한 칸만 확인하고 나머지 칸이 왜 따라 나오는지 말하지 않은 채 X_{12}와 X_{23}이 독립이라고 주장하는 것. 두 값만 갖는 변수에서는 실제로 따라 나오지만, 그 사실을 밝혀야 한다.
- (iii)에서 선수를 공유하는 쌍들은 공통인 수가 짝수이든 홀수이든 $+\frac{1}{36}$을 기여한다는 점을 잊는 것. 곱에서 음의 부호는 상쇄된다.
- 겹치는 쌍들의 쌍을 잘못 세는 것. 예를 들어 (순서 있는) $n(n-1)(n-2)$를 쓰고 거기에 또 2를 곱하는 경우.

STEP 너머

소문항 (i)은 **쌍별** 독립(pairwise independent)이지만 **상호** 독립(mutually independent)은 아닌 변수의 고전적인 예이다. $X_{12} = X_{23} = 1$이면 $X_{13} = 1$이 강제되기 때문이다. 분산 계산에는 쌍별 독립만 있으면 충분하며, 이 아이디어는 무작위 알고리즘(randomised algorithms)과 생일 문제(birthday problem)에 쓰인다. 소문항 (iii)은 항들이 종속일 때 합의 공분산을 다루는 일반적인 방법을 보여 준다. 두 내용 모두 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에서 전개된다.

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.