풀이

STEP III 2002, 문제 14

확률과 통계 • 분산과 공분산

> **STEP III 2002, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/02-s3-q14

핵심 아이디어

분산 항등식은 분산(variance)의 정의를 기댓값의 선형성(linearity of expectation)으로 전개한 것일 뿐이다. 체스 부분은 보기보다 영리하다. 우리는 $\mathrm{E}(WL)$을 직접 계산하지 않는다. 그 대신 이항분포(binomial distribution)로부터 $W+L$의 분산을 알고, 역시 이항분포로부터 $\mathrm{Var}(W)$와 $\mathrm{Var}(L)$을 알며, 방금 증명한 항등식이 공분산(covariance)을 공짜로 건네준다.

풀이

분산 항등식

$\mu_X = \mathrm{E}(X)$, $\mu_Y = \mathrm{E}(Y)$라 하자. 다음 정의

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu_X)^2], \qquad \mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)],$$

와 기댓값의 선형성 $\mathrm{E}(aU+bV) = a\,\mathrm{E}(U) + b\,\mathrm{E}(V)$를 사용한다. 이산 확률변수(discrete random variable)의 경우 이 선형성은 합 $\sum(au+bv)\,\mathrm{P}(U=u, V=v)$를 두 개의 합으로 나누고 각 합에서 다른 변수에 대해 합하여 없애면 얻어진다.

선형성에 의해 $\mathrm{E}(X+Y) = \mu_X + \mu_Y$이므로

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}[((X-\mu_X)+(Y-\mu_Y))^2] \\
&= \mathrm{E}[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] \\
&= \mathrm{E}[(X-\mu_X)^2] + \mathrm{E}[(Y-\mu_Y)^2] + 2\,\mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y),
\end{aligned}$$

이다. 셋째 줄에서 선형성을 다시 사용하였다. 이것이 구하려는 결과이다. 독립성(independence)이 전혀 필요하지 않았다는 점에 주목하라. 이 항등식은 임의의 X와 Y에 대해 성립한다.

소문항 (i)

모든 경기는 승, 패, 무승부 중 하나이므로 어떤 일이 일어나든 $W + L + D = m$이다. 항상 같은 값을 갖는 확률변수는 그 평균과 같으므로, 분산은 $\mathrm{E}[(m-m)^2] = 0$이다.

소문항 (ii)

승부가 난 경기, 즉 이기거나 진 경기를 **결판난** 경기라 부르자. 각 경기가 결판날 확률은 $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$이고, 경기들은 서로 독립이며, 경기 수는 m으로 고정되어 있다. 따라서 결판난 경기의 수 $W + L$은 성공 확률이 같은 m번의 독립 시행에서 성공 횟수를 센 것이다.

$$W + L \sim \mathrm{B}(m, \tfrac{2}{3}).$$

상관계수

정확히 같은 방식으로, W는 각각 확률 $\frac{1}{3}$로 이기는 m번의 독립 시행에서 승리 횟수를 센 것이므로 $W \sim \mathrm{B}(m, \frac{1}{3})$이고, 마찬가지로 $L \sim \mathrm{B}(m, \frac{1}{3})$이다. 이항분포의 분산 $\mathrm{Var} = np(1-p)$를 사용하면

$$\mathrm{Var}(W) = \mathrm{Var}(L) = m \cdot \tfrac{1}{3} \cdot \tfrac{2}{3} = \frac{2m}{9}, \qquad \mathrm{Var}(W+L) = m \cdot \tfrac{2}{3} \cdot \tfrac{1}{3} = \frac{2m}{9}.$$

? **생각해 보기** $\mathrm{E}(WL)$을 계산하지 않고 어떻게 $\mathrm{Cov}(W, L)$을 구할 수 있을까?

항등식을 거꾸로 사용한다. $X = W, Y = L$로 두면

$$\frac{2m}{9} = \frac{2m}{9} + \frac{2m}{9} + 2\,\mathrm{Cov}(W, L), \qquad \text{따라서} \qquad \mathrm{Cov}(W, L) = -\frac{m}{9}.$$

그러므로 $m \geq 1$일 때(분산이 0이 아니도록),

$$\frac{\mathrm{Cov}(W, L)}{\sqrt{\mathrm{Var}(W)\,\mathrm{Var}(L)}} = \frac{-m/9}{2m/9} = -\frac{1}{2}.$$

이 답은 m에 의존하지 않는다. 그리고 당연히 음수이다. 경기 수가 고정되어 있으므로 승리가 많을수록 패배가 들어갈 자리가 줄어든다.

소문항 (i)은 두 번째 경로를 제공한다. 항등식을 두 번 적용하면 $\mathrm{Var}(W + L + D)$는 세 분산의 합에 세 쌍별 공분산의 합의 두 배를 더한 것이다. 대칭성에 의해 세 공분산은 모두 어떤 값 c로 같으므로 $0 = 3 \cdot \frac{2m}{9} + 6c$이고, 다시 $c = -\frac{m}{9}$를 얻는다.

검산

- $m = 1$로 두자. 그러면 W와 L은 $\mathrm{E}(W) = \mathrm{E}(L) = \frac{1}{3}$인 지시변수(indicator)이고 항상 $WL = 0$이므로 $\mathrm{Cov}(W, L) = 0 - \frac{1}{9} = -\frac{1}{9}$이며, 이는 $-m/9$와 일치한다.
- 두 경로(소문항 (ii)를 통한 것과 소문항 (i)을 통한 것)의 결과가 일치한다.
- $m = 1, \ldots, 7$에 대해 3^m가지 결과를 모두 정확히 나열하여 계산하면 매번 $\mathrm{Var}(W) = 2m/9$, $\mathrm{Cov}(W, L) = -m/9$이고 상관계수(correlation)는 정확히 $-\frac{1}{2}$이다.

흔한 실수

- **경기들**이 독립이므로 W와 L도 독립이라고 가정하는 것. 그렇지 않다. 이들은 D와 함께 고정된 합을 공유한다.
- $\mathrm{Var}(W + L) = \mathrm{Var}(W) + \mathrm{Var}(L) = 4m/9$로 두는 것. 이는 공분산 항을 빠뜨린 것이며 이항분포로 얻은 답 $2m/9$와 모순된다.

STEP 너머

(W, L, D)는 **다항분포**(multinomial distribution)를 따르는데, 이는 시행마다 결과가 두 가지보다 많은 경우로 이항분포를 자연스럽게 일반화한 것이다. 일반적으로 각 결과의 확률이 $p_1, \ldots, p_k$이면, 횟수 N_i와 N_j의 공분산은 $-mp_ip_j$이고 상관계수는 $-\sqrt{p_ip_j/\big((1-p_i)(1-p_j)\big)}$이며, 여기서는 $-\frac{1}{2}$가 된다. 공분산 행렬(covariance matrix)과 공분산의 쌍선형성(bilinearity)은 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문), 그리고 Blitzstein과 Hwang의 *Introduction to Probability*(확률 입문)에서 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.