풀이

STEP I 2018, 문제 12

확률과 통계 • 복원추출

> **STEP I 2018, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/18-s1-q12

핵심 아이디어

앞면이 나온 횟수를 각 던지기마다 하나씩, 두 지시변수(indicator variable)의 합으로 나타낸다. 합의 기댓값(expectation)은 던지기들이 독립(independent)이든 아니든 기댓값의 합이므로, 평균은 쉽게 구해진다. 분산(variance)은 단 하나의 수, 곧 두 던지기 모두 앞면이 나올 확률에 달려 있으며, 바로 이 지점에서 복원추출(drawing with replacement)과 비복원추출(drawing without replacement)이 갈린다.

풀이

풀이 전체에서 $p = \frac{1}{3}(p_1 + p_2 + p_3)$이고

$$q = p_1p_2 + p_2p_3 + p_3p_1.$$

라 하자. 각 동전이 선택될 가능성은 똑같고, 일단 동전이 선택되면 그 던지기들은 서로 독립이라고 가정한다.

소문항 (i)

동전 k가 선택되는 사건을 C_k라 하면 $\mathrm{P}(C_k) = \frac{1}{3}$이다. 전확률 법칙(law of total probability)에 의해

$$\mathrm{P}(\text{앞면}) = \sum_{k=1}^{3} \mathrm{P}(C_k)\,\mathrm{P}(\text{앞면} \mid C_k) = \tfrac{1}{3}(p_1 + p_2 + p_3) = p.$$

소문항 (ii)

첫 번째와 두 번째 던지기에서 앞면이 나오는 것의 지시변수를 각각 X_1, X_2라 하면 $N_1 = X_1 + X_2$이다. 첫 번째 동전을 되돌려 놓으므로 두 번째 던지기는 첫 번째와 완전히 같은 상황에서 일어나고, 소문항 (i)에 의해 $\mathrm{P}(X_1 = 1) = \mathrm{P}(X_2 = 1) = p$이다. 따라서

$$\mathrm{E}(N_1) = \mathrm{E}(X_1) + \mathrm{E}(X_2) = 2p.$$

분산을 구하려면 $\mathrm{P}(X_1 = 1,\, X_2 = 1)$이 필요하다. 두 번의 선택은 독립이고, 동전의 순서쌍 (j,k) 9개는 각각 확률 $\frac{1}{9}$을 가지며, 순서쌍이 주어졌을 때 두 번 모두 앞면이 나올 확률은 $p_j p_k$이다. 그러므로

$$\mathrm{P}(X_1 = 1,\, X_2 = 1) = \frac{1}{9}\sum_{j=1}^{3}\sum_{k=1}^{3} p_j p_k = \left(\frac{p_1 + p_2 + p_3}{3}\right)^2 = p^2 = \mathrm{P}(X_1 = 1)\,\mathrm{P}(X_2 = 1).$$

다른 결과들에 대해서도 같은 방식으로 곱으로 분해되므로 X_1과 X_2는 독립이고 $N_1 \sim \mathrm{B}(2,p)$이다. 직접 계산하면, $X_i^2 = X_i$이므로

$$\begin{aligned}\mathrm{E}(N_1^2) &= \mathrm{E}(X_1) + \mathrm{E}(X_2) + 2\,\mathrm{E}(X_1X_2) = 2p + 2p^2,\\ \mathrm{Var}(N_1) &= 2p + 2p^2 - (2p)^2 = 2p(1-p),\end{aligned}$$

이 되어 요구한 결과를 얻는다.

소문항 (iii)

이제 두 동전을 비복원으로 하나씩 차례로 고르고, 첫 번째와 두 번째로 고른 동전에서 앞면이 나오는 것의 지시변수를 Y_1, Y_2라 하면 $N_2 = Y_1 + Y_2$이다. 각 동전이 첫 번째로 뽑힐 가능성은 똑같고, 대칭성(symmetry)에 의해 두 번째로 뽑힐 가능성도 똑같으므로, 소문항 (i)과 마찬가지로 $\mathrm{P}(Y_1 = 1) = \mathrm{P}(Y_2 = 1) = p$이고

$$\mathrm{E}(N_2) = 2p.$$

생각해 보기 소문항 (ii)에서 했던 P(둘 다 앞면)의 계산을 다시 쓸 수 있을까?

그렇지 않다. 이제는 같은 동전을 두 번 뽑을 수 없기 때문이다. $j \neq k$인 순서쌍 (j,k) 6개가 같은 확률로 나오고, 순서 없는 쌍 $\{j,k\}$ 각각은 두 번씩 나타나므로

$$\mathrm{P}(Y_1 = 1,\, Y_2 = 1) = \frac{1}{6}\sum_{j\neq k} p_j p_k = \frac{2q}{6} = \frac{q}{3}.$$

이 값은 일반적으로 p^2이 아니므로 Y_1과 Y_2는 독립이 아니다. 다시 $Y_i^2 = Y_i$를 이용하면

$$\begin{aligned}\mathrm{E}(N_2^2) &= \mathrm{E}(Y_1) + \mathrm{E}(Y_2) + 2\,\mathrm{E}(Y_1Y_2) = 2p + \tfrac{2}{3}q,\\ \mathrm{Var}(N_2) &= 2p + \tfrac{2}{3}(p_1p_2 + p_2p_3 + p_3p_1) - 4p^2.\end{aligned}$$

소문항 (iv)

두 식을 빼면

$$\mathrm{Var}(N_1) - \mathrm{Var}(N_2) = (2p - 2p^2) - (2p + \tfrac{2}{3}q - 4p^2) = 2p^2 - \tfrac{2}{3}q.$$

그런데 $9p^2 = (p_1 + p_2 + p_3)^2 = p_1^2 + p_2^2 + p_3^2 + 2q$이므로

$$\begin{aligned} 2p^2 - \tfrac{2}{3}q &= \tfrac{2}{9}(p_1^2 + p_2^2 + p_3^2 + 2q) - \tfrac{6}{9}q = \tfrac{2}{9}(p_1^2 + p_2^2 + p_3^2 - p_1p_2 - p_2p_3 - p_3p_1) \\ &= \tfrac{1}{9}\Big[(p_1 - p_2)^2 + (p_2 - p_3)^2 + (p_3 - p_1)^2\Big], \end{aligned}$$

이다. 마지막 단계는 세 제곱을 전개하면 $2(p_1^2 + p_2^2 + p_3^2) - 2q$가 되기 때문이다.

실수의 제곱의 합은 음이 아니므로 $\mathrm{Var}(N_2) \le \mathrm{Var}(N_1)$이다. 등호는 세 제곱이 모두 0일 때, 곧 $p_1 = p_2 = p_3$일 때에만 성립한다. 이로써 증명이 끝난다.

부산물로 소문항 (iii)의 답을 깔끔한 꼴로 얻는다.

$$\mathrm{Var}(N_2) = 2p(1-p) - \tfrac{1}{9}\Big[(p_1 - p_2)^2 + (p_2 - p_3)^2 + (p_3 - p_1)^2\Big].$$

검산

- $p_1 = p_2 = p_3 = p$이면 모든 동전이 똑같으므로 동전을 어떻게 고르든 상관이 없다. 두 횟수 모두 $\mathrm{B}(2, p)$이고 분산이 일치하므로, 등호가 성립하는 경우와 맞아떨어진다.
- $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$으로 두면 $p = \frac{1}{3}$, $q = \frac{11}{36}$이다. 그러면 $\mathrm{Var}(N_1) = \frac{4}{9}$이고 $\mathrm{Var}(N_2) = \frac{2}{3} + \frac{11}{54} - \frac{4}{9} = \frac{23}{54}$이므로 $\frac{1}{54}$만큼 작으며, 이는 $\frac{1}{9}\left(\frac{1}{36} + \frac{1}{36} + \frac{1}{9}\right) = \frac{1}{54}$와 일치한다. 모든 결과를 손으로 나열해 보아도 같은 값을 얻는다.
- 극단적인 경우 $p_1 = p_2 = 1$, $p_3 = 0$: 소문항 (iii)에서 동전 1과 2가 함께 선택되는 경우를 제외하면, 선택된 두 동전 중 정확히 하나가 반드시 앞면이 나오는 동전이다. 따라서 N_2는 확률 $\frac{1}{3}$로 2이고 그 밖에는 1이므로 분산은 $\frac{2}{9}$이다. 공식으로 계산하면 $2p + \frac{2}{3}q - 4p^2 = \frac{4}{3} + \frac{2}{3} - \frac{16}{9} = \frac{2}{9}$이다.

흔한 실수

- 소문항 (iii)의 두 던지기를 독립으로 취급하여 다시 $2p(1-p)$를 얻는 것. 소문항 (iv)의 요점은 바로 그것들이 독립이 아니라는 데 있다.
- $\mathrm{Var}(N_1)$을 각 동전의 분산 $2p_k(1-p_k)$의 평균으로 쓰는 것. 이는 동전을 무작위로 고르는 데서 생기는 추가적인 변동을 무시하며, 소문항 (ii)에서는 던질 때마다 다른 동전이 쓰일 수도 있다.

STEP 너머

소문항 (ii)와 (iii)의 차이는 복원 표본추출(sampling)과 비복원 표본추출의 차이이다. 비복원추출을 하면 뽑힌 것들 사이에 음의 상관관계(negative correlation)가 생기고, 이 때문에 분산이 줄어든다. 같은 효과가 표본조사에서 "유한 모집단 수정(finite population correction)"을 낳으며, 초기하분포(hypergeometric distribution)가 평균이 같은 이항분포(binomial distribution)보다 분산이 작은 이유이기도 하다. 소문항 (iv)의 대수 계산은 항등식(identity) $\sum p_k^2 - \sum_{j<k} p_j p_k = \frac{1}{2}\sum_{j<k}(p_j - p_k)^2$이며, 분산이 결코 음이 아니라는 사실과 가까운 사촌 관계에 있다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 지시변수, 공분산(covariance), 초기하분포를 바로 이러한 관점에서 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.