풀이

STEP III 2007, 문제 12

확률과 통계 • 무작위 합

STEP III 2007, 12번 문제

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

step.maths.org/questions/07-s3-q12

핵심 아이디어

Y는 항의 개수가 **확률적으로 정해지는** 합이므로, 고정된 개수의 확률변수의 합에 대한 규칙을 곧바로 쓸 수는 없다. 요령은 N에 대해 조건을 거는(conditioning) 것이다. $N = k$임을 알고 나면 Y는 평균과 분산을 알고 있는 k개의 독립인 확률변수의 평범한 합이 된다. 그다음 N의 분포에 대해 평균을 내면 되는데, 첫 부분에서 $\mathrm{E}(N)$과 $\mathrm{E}(N^2)$을 구하라고 하는 것도 이 때문이다.

풀이

N의 분포

수가 균등하게 무작위로 선택된다고 하면 $\mathrm{P}(N = k) = \frac{1}{2n-1}$ $(k = 1, 2, \ldots, 2n-1)$이다. $\sum_{k=1}^{m} k = \frac{1}{2}m(m+1)$과 $\sum_{k=1}^{m} k^2 = \frac{1}{6}m(m+1)(2m+1)$에 $m = 2n-1$을 대입하면 다음을 얻는다.

$$\mathrm{E}(N) = \frac{1}{2n-1} \cdot \frac{(2n-1)(2n)}{2} = n,$$
$$\mathrm{E}(N^2) = \frac{1}{2n-1} \cdot \frac{(2n-1)(2n)(4n-1)}{6} = \frac{n(4n-1)}{3}.$$

(평균이 n이라는 것은 대칭성으로도 분명하다. n은 $1, \ldots, 2n-1$의 한가운데 값이다.) 나중에 쓰기 위해 다음도 구해 둔다.

$$\mathrm{Var}(N) = \frac{n(4n-1)}{3} - n^2 = \frac{n^2 - n}{3} = \frac{n(n-1)}{3}.$$

N에 대한 조건부

k를 고정하고 $N = k$라고 하자. X_i들은 N과 독립이므로 $N = k$라는 조건을 걸어도 그 분포는 바뀌지 않는다. 따라서 $N = k$일 때 $Y = X_1 + \cdots + X_k$이고, 이는 각각 평균이 μ, 분산이 σ^2인 k개의 독립인 확률변수의 합이다. 그러므로

$$\mathrm{E}(Y \mid N = k) = k\mu, \qquad \mathrm{Var}(Y \mid N = k) = k\sigma^2,$$

이고, 두 번째 식은 X_i들의 독립성(분산이 더해진다)을 이용한 것이다. 따라서

$$\mathrm{E}(Y^2 \mid N = k) = \mathrm{Var}(Y \mid N = k) + (\mathrm{E}(Y \mid N = k))^2 = k\sigma^2 + k^2\mu^2.$$

> **생각해 보기** 조건부 기댓값(conditional expectation)에서 어떻게 보통의 기댓값으로 되돌아올 수 있을까?

전체 기댓값의 법칙(law of total expectation, 분할 정리)에 의해 $\mathrm{E}(g(Y,N)) = \sum_k \mathrm{P}(N = k)\,\mathrm{E}(g(Y,N) \mid N = k)$이다. 각 항 안에서 N은 그냥 수 k이다.

Y의 평균

$$\mathrm{E}(Y) = \sum_k \mathrm{P}(N = k)\, k\mu = \mu\,\mathrm{E}(N) = n\mu,$$

이므로 요구된 결과가 성립한다.

공분산

$N = k$일 때 $YN = kY$이므로 $\mathrm{E}(YN \mid N = k) = k \cdot k\mu = k^2\mu$이다. 따라서 $\mathrm{E}(YN) = \mu\,\mathrm{E}(N^2)$이고

$$\begin{aligned}\mathrm{Cov}(Y,N) &= \mathrm{E}(YN) - \mathrm{E}(Y)\,\mathrm{E}(N) = \mu\,\mathrm{E}(N^2) - n\mu \cdot n \\ &= \mu\,\mathrm{Var}(N) = \tfrac{1}{3}n(n-1)\mu,\end{aligned}$$

이므로 요구된 결과를 얻는다.

Y의 분산

$\mathrm{E}(Y^2 \mid N = k) = k\sigma^2 + k^2\mu^2$를 N에 대해 평균 내면

$$\mathrm{E}(Y^2) = \sigma^2\,\mathrm{E}(N) + \mu^2\,\mathrm{E}(N^2) = n\sigma^2 + \frac{n(4n-1)}{3}\mu^2.$$

따라서

$$\mathrm{Var}(Y) = n\sigma^2 + \frac{n(4n-1)}{3}\mu^2 - n^2\mu^2 = n\sigma^2 + \frac{n(n-1)}{3}\mu^2.$$

두 항에는 자연스러운 의미가 있다. $n\sigma^2$은 시행 횟수가 고정되어 있을 때 Y가 퍼진 정도의 평균이고, $\frac{1}{3}n(n-1)\mu^2 = \mu^2\,\mathrm{Var}(N)$은 시행이 몇 번 일어날지 모르기 때문에 생기는 추가적인 퍼짐이다.

검산

- $n = 1$이면 항상 $N = 1$이고, 공식들은 $\mathrm{E}(Y) = \mu$, $\mathrm{Cov}(Y, N) = 0$, $\mathrm{Var}(Y) = \sigma^2$으로 줄어든다. X_1 하나뿐일 때 당연히 그래야 한다.
- $\mu = 0$이면 추가 항이 사라진다. 항의 개수가 더 이상 평균에 영향을 주지 않으므로, 퍼짐을 더하거나 Y와 상관관계를 가질 수 없다.
- $n = 2$로 두고 각 X_i가 확률 $\frac{1}{2}$로 0 또는 1이라고 하자. 그러면 $N \in \{1, 2, 3\}$이고 $Y \mid N \sim \mathrm{B}(N, \frac{1}{2})$이며, 공식은 $\mathrm{Var}(Y) = 2 \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{4} = \frac{2}{3}$을 준다. 이 값과 모든 결과를 $n = 1, \ldots, 6$ 및 여러 가지 X_i의 분포에 대해 (N, Y)의 결합분포(joint distribution)를 정확히 열거하여 확인하였다.

흔한 실수

- N 자체가 변한다는 것을 잊고 $\mathrm{Var}(Y) = \mathrm{E}(N)\sigma^2$만 쓰는 것. 두 번째 항이 바로 핵심이다.
- Y와 N을 독립으로 취급하여 $\mathrm{Cov}(Y, N) = 0$을 얻는 것. $\mu > 0$이면 시행이 많을수록 합이 커지므로, 둘은 양의 상관관계를 가진다.

STEP 너머

Y는 **확률적 합**(random sum, 복합분포(compound distribution))이고, 분산에 대한 결과는 전체 분산의 법칙(law of total variance) $\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y \mid N)) + \mathrm{Var}(\mathrm{E}(Y \mid N))$의 한 경우로, 여기서는 $\sigma^2\mathrm{E}(N) + \mu^2\mathrm{Var}(N)$이 된다. 확률적 합은 보험금 청구와 분기 과정(branching process)을 모형화하며, 확률생성함수(probability generating function)를 이용하면 우아하게 다룰 수 있다: $G_Y(s) = G_N(G_X(s))$. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 확률적 합과 분기 과정을 이런 방식으로 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.