풀이

STEP III 2006, 문제 14

확률과 통계 • 공분산과 독립

> **STEP III 2006, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/06-s3-q14

핵심 아이디어

이 문제는 모두 두 가지 사실로 풀린다. 기댓값(expectation)은 선형(linear)이고, 독립(independent)인 확률변수에 대해서는 곱의 기댓값이 기댓값의 곱과 같다는 것이다. 분산(variance)은 $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2$에서 얻으므로, 결국 기댓값만 구하면 된다. 두 확률변수가 독립이 아님을 보이려면 독립성에서 따라 나오는 결과 하나가 성립하지 않음을 찾으면 충분하며, 가장 확인하기 쉬운 것은 $\mathrm{E}(UV) = \mathrm{E}(U)\,\mathrm{E}(V)$이다.

풀이

두 가지 사실

임의의 확률변수(random variable)와 상수 a, b에 대하여

$$\mathrm{E}(aX_1 + bX_2) = a\,\mathrm{E}(X_1) + b\,\mathrm{E}(X_2),$$

이고, X_1과 X_2가 독립이면

$$\mathrm{E}(X_1X_2) = \mathrm{E}(X_1)\,\mathrm{E}(X_2).$$

이다. 두 번째 사실은 확률변수의 함수에도 적용된다. X_1, X_2가 독립이면 X_1^2과 X_2, X_1과 X_2^2, 그리고 X_1^2과 X_2^2도 각각 독립이다. 또한 $\mathrm{E}(X_i^2) = \sigma_i^2 + \mu_i^2$을 사용할 것인데, 이것은 $\mathrm{Var}(X_i) = \mathrm{E}(X_i^2) - \mu_i^2$을 정리한 것에 지나지 않는다.

둘레와 넓이

둘레(perimeter)는 $P = 2X_1 + 2X_2$이고 넓이(area)는 $A = X_1X_2$이다. 선형성과 독립성에 의하여

$$\mathrm{E}(P) = 2(\mu_1 + \mu_2), \qquad \mathrm{E}(A) = \mu_1\mu_2.$$

이다. P의 분산은 $\mathrm{E}(P^2)$에서 출발하여 구한다.

$$\begin{aligned}\mathrm{E}(P^2) &= 4\,\mathrm{E}(X_1^2 + 2X_1X_2 + X_2^2)\\ &= 4(\sigma_1^2 + \mu_1^2 + 2\mu_1\mu_2 + \sigma_2^2 + \mu_2^2)\\ &= 4(\sigma_1^2 + \sigma_2^2) + 4(\mu_1 + \mu_2)^2,\end{aligned}$$

여기서 선형성과 $\mathrm{E}(X_1X_2) = \mu_1\mu_2$를 사용하였다. $[\mathrm{E}(P)]^2 = 4(\mu_1 + \mu_2)^2$을 빼면

$$\mathrm{Var}(P) = 4(\sigma_1^2 + \sigma_2^2), \qquad \text{따라서} P\text{의 표준편차는} 2\sqrt{\sigma_1^2 + \sigma_2^2}.$$

이다. 즉 P의 표준편차(standard deviation)가 구해진다. A의 경우, X_1^2과 X_2^2이 독립이므로

$$\mathrm{E}(A^2) = \mathrm{E}(X_1^2)\,\mathrm{E}(X_2^2) = (\sigma_1^2 + \mu_1^2)(\sigma_2^2 + \mu_2^2),$$

이고, 여기서 $\mu_1^2\mu_2^2$을 빼면

$$\mathrm{Var}(A) = \sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2,$$

이 남는다. 따라서 A의 표준편차는

$$\sqrt{\sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}.$$

이다.

P와 A는 독립이 아니다

P와 A가 독립이라면 $\mathrm{E}(PA) = \mathrm{E}(P)\,\mathrm{E}(A)$가 성립해야 한다. 그런데 $PA = 2X_1^2X_2 + 2X_1X_2^2$이므로, 선형성과 X_1, X_2의 독립성에 의하여

$$\begin{aligned}\mathrm{E}(PA) &= 2(\sigma_1^2 + \mu_1^2)\mu_2 + 2\mu_1(\sigma_2^2 + \mu_2^2)\\ &= 2(\mu_1 + \mu_2)\mu_1\mu_2 + 2(\mu_2\sigma_1^2 + \mu_1\sigma_2^2)\\ &= \mathrm{E}(P)\,\mathrm{E}(A) + 2(\mu_2\sigma_1^2 + \mu_1\sigma_2^2).\end{aligned}$$

이다. 가로와 세로의 길이는 양수이므로 $\mu_1, \mu_2 > 0$이고, $\sigma_1, \sigma_2 \neq 0$이므로 추가된 항 $2(\mu_2\sigma_1^2 + \mu_1\sigma_2^2)$은 순양수이다. 따라서 $\mathrm{E}(PA) \neq \mathrm{E}(P)\,\mathrm{E}(A)$이고, P와 A는 독립이 아니다.

확률변수 $Z = P - \alpha A$

이번에도 Z와 A가 독립이라면 $\mathrm{E}(ZA) = \mathrm{E}(Z)\,\mathrm{E}(A)$가 성립해야 한다. 선형성에 의하여

$$\begin{aligned}\mathrm{E}(ZA) - \mathrm{E}(Z)\,\mathrm{E}(A) &= [\mathrm{E}(PA) - \mathrm{E}(P)\,\mathrm{E}(A)] - \alpha[\mathrm{E}(A^2) - [\mathrm{E}(A)]^2] \\ &= 2(\mu_2\sigma_1^2 + \mu_1\sigma_2^2) - \alpha(\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_1^2\sigma_2^2),\end{aligned}$$

이며, 여기서 앞의 두 결과를 사용하였다. α에 곱해진 괄호는 $\mathrm{Var}(A)$이고, 이는 양수이다($\sigma_1^2\sigma_2^2 > 0$을 포함하므로). 따라서 이 식은 정확히 하나의 α 값에서만 0이 된다.

$$\alpha_0 = \frac{2(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)}{\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_1^2\sigma_2^2}.$$

$\alpha \neq \alpha_0$이면 $\mathrm{E}(ZA) \neq \mathrm{E}(Z)\,\mathrm{E}(A)$이므로, 요구한 대로 Z와 A는 독립이 아니다.

? **생각해 보기** $\alpha = \alpha_0$이면 Z와 A가 독립인가?

아니다. 공분산(covariance)이 0인 것은 독립성의 결과일 뿐, 독립성을 보장하지는 않는다. $\alpha = \alpha_0$에서는 우리의 판정법이 아무 정보도 주지 않으므로 다른 논증이 필요하다. 마지막 부분에서는 하나의 구체적인 경우에 그때에도 독립성이 성립하지 않음을 보인다.

마지막 부분

여기서는 $\mu_1 = \mu_2 = 2$이고 $\sigma_1^2 = \sigma_2^2 = 1$이므로

$$\alpha_0 = \frac{2(2+2)}{4+4+1} = \frac{8}{9}.$$

이다. 앞 부분에 의하여 $\alpha \neq \frac{8}{9}$인 모든 α에 대해 Z와 A는 독립이 아니다. 모든 α를 한꺼번에 다루는 논증을 위해, 일어날 가능성이 같은 네 가지 결과를 나열해 보자.

(X_1, X_2)	A	P	$Z = P - \alpha A$
$(1,1)$	1	4	$4 - \alpha$
$(1,3)$, $(3,1)$	3	8	$8 - 3\alpha$
$(3,3)$	9	12	$12 - 9\alpha$

핵심 관찰은 P, 따라서 Z가 A에 의해 완전히 결정된다는 것이다. $z = 4 - \alpha$라 하자. $A = 1$임을 알면 $Z = z$일 수밖에 없으므로

$$\mathrm{P}(Z = z,\ A = 1) = \mathrm{P}(A = 1) = \tfrac{1}{4}.$$

이다. 독립이라면 이 값이 $\mathrm{P}(Z = z)\,\mathrm{P}(A = 1)$과 같아야 하고, 즉 $\mathrm{P}(Z = z) = 1$이어야 하므로, Z는 모든 결과에서 $4 - \alpha$라는 값을 가져야 한다. 그러려면 $8 - 3\alpha = 4 - \alpha$에서 $\alpha = 2$이

어야 하고, 동시에 $12 - 9\alpha = 4 - \alpha$에서 $\alpha = 1$이어야 한다. 이 둘은 동시에 성립할 수 없으므로 $\mathrm{P}(Z = z) < 1$이고, 독립성은 성립하지 않는다.

따라서 어떤 α 값에 대해서도 Z와 A는 독립이 아니다. 예를 들어 $\alpha = \frac{8}{9}$일 때 Z의 값은 $\frac{28}{9}$, $\frac{16}{3}$, 4이고, $\mathrm{P}(Z = \frac{28}{9}, A = 1) = \frac{1}{4}$인 반면 $\mathrm{P}(Z = \frac{28}{9})\,\mathrm{P}(A = 1) = \frac{1}{16}$이다.

검산

- 변동이 없으면($\sigma_1 = \sigma_2 = 0$) 두 표준편차가 모두 0이 되며, 이는 당연한 결과이다.
- 마지막 부분에서 공식에 의하면 $\mathrm{Var}(A) = 1 + 4 + 4 = 9$이다. 직접 계산하면 A는 $1, 3, 9$의 값을 각각 확률 $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$로 가지므로 $\mathrm{E}(A) = 4$, $\mathrm{E}(A^2) = 25$이고 분산은 9이다.
- 일반적인 평균, 분산, 공분산 공식을 무작위로 만든 200개의 이산 분포에 대해 분수를 사용하여 정확하게 확인하였다. 또한 α_0에서 $\mathrm{E}(ZA) = \mathrm{E}(Z)\mathrm{E}(A)$가 성립함, 마지막 부분에서 $\alpha_0 = \frac{8}{9}$임, 그리고 그 경우 $\alpha = \frac{8}{9}$와 다른 500개의 α 값에 대해 Z와 A가 독립이 아님을 확인하였다.

흔한 실수

- $\mathrm{E}(X_1^2) = \mu_1^2$이라고 쓰는 것. 분산 항 σ_1^2이 바로 $\mathrm{Var}(A)$와 공분산을 0이 아니게 만드는 요소이다.
- $\alpha = \alpha_0$일 때 Z와 A가 독립이라고 결론 내리는 것. 공분산이 0이라고 해서 독립인 것은 아니다.
- $\mu_2\sigma_1^2 + \mu_1\sigma_2^2 \neq 0$인 이유를 밝히지 않는 것. 항들이 상쇄될 가능성을 배제하는 것은 길이가 양수라는 사실이다.

STEP 너머

$\mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V)$라는 양이 공분산이며, 특별한 값 α_0은 $\mathrm{Cov}(P, A)/\mathrm{Var}(A)$, 즉 A에 대한 P의 최소제곱 회귀(least-squares regression) 직선의 기울기이다. $\alpha = \alpha_0$으로 택하면 잔차(residual) Z는 A와 무상관(uncorrelated)이 되지만, A와 독립이 되지는 않는다. 결합정규분포(jointly normal)를 따르는 확률변수에서는 두 개념이 일치하며, 이것이 정규분포가 그토록 편리한 이유 중 하나이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 공분산, 상관(correlation), 독립성을 꼼꼼하게 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.