풀이

STEP III 2010, 문제 13

확률과 통계 • 선형결합의 상관관계

> **STEP III 2010, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/10-s3-q13

이 문제는 *Beyond STEP*의 13.7 절 **상관관계 만들어 내기**에서 단계별 안내와 함께 자세히 다룬다: https://beyondstep.keithmcnulty.org/ko/ch13/correlation.html

핵심 아이디어

여기에 나오는 모든 것은 평균이 0이고 분산이 1인 독립(independent) 확률변수들의 일차결합(linear combination)이므로, 모든 기댓값, 분산, 공분산(covariance)을 계수만 보고 바로 읽어 낼 수 있다. 즉 $U = \sum u_k Z_k$이고 $V = \sum v_k Z_k$이면 $\mathrm{Cov}(U, V) = \sum u_k v_k$이다. 이 문제는 사실 변수를 하나씩 추가하면서, 주어진 임의의 상관계수(correlation)를 갖는 세 개의 표준화된 변수를 만드는 것이며, 이것은 모습을 바꾼 그람-슈미트 과정(Gram–Schmidt process)이다. 새로운 조건 하나가 새로운 계수 하나를 결정한다.

풀이

풀이 전체에서 공분산의 쌍선형성(bilinearity), $\mathrm{Var}(Z_k) = 1$, 그리고 $j \neq k$일 때 $\mathrm{Cov}(Z_j, Z_k) = 0$(이것은 Z_k들이 독립이므로 성립한다)을 사용한다. 특히 상수 u_k, v_k에 대하여

$$\mathrm{Cov}\Big(\sum_k u_k Z_k,\ \sum_k v_k Z_k\Big) = \sum_{j,k} u_j v_k \,\mathrm{Cov}(Z_j, Z_k) = \sum_k u_k v_k.$$

Z_1과 Z_2의 상관계수

독립성에 의해 $\mathrm{Cov}(Z_1, Z_2) = \mathrm{E}(Z_1 Z_2) - \mathrm{E}(Z_1)\mathrm{E}(Z_2) = 0$이므로 $\mathrm{Corr}(Z_1, Z_2) = 0$이다.

변수 Y_2

$s = \sqrt{1-\rho_{12}^2}$로 놓자. $-1 < \rho_{12} < 1$이므로 s는 양의 실수이다. 그러면 $Y_2 = \rho_{12}Z_1 + sZ_2$이고, 기댓값의 선형성(linearity of expectation)에 의해

$$\mathrm{E}(Y_2) = \rho_{12}\mathrm{E}(Z_1) + s\,\mathrm{E}(Z_2) = 0.$$

위의 공분산 규칙에 의해

$$\mathrm{Var}(Y_2) = \rho_{12}^2 + s^2 = \rho_{12}^2 + (1-\rho_{12}^2) = 1, \qquad \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Z_1, \rho_{12}Z_1 + sZ_2) = \rho_{12}.$$

$\mathrm{Var}(Y_1) = \mathrm{Var}(Y_2) = 1$이므로 정의에서 분모는 1이고,

$$\mathrm{Corr}(Y_1, Y_2) = \rho_{12}.$$

변수 Y_3

조건 $\mathrm{E}(Y_3) = 0$은 모든 a, b, c에 대하여 자동으로 성립하므로 아무 정보도 주지 않는다. 나머지 세 조건은 각각 방정식 하나씩을 준다. Y_1, Y_2, Y_3의 분산이 모두 1이므로 각 상관계수는 대응하는 공분산과 같다.

Y_1**과의 상관계수.** $\mathrm{Cov}(Z_1,\ aZ_1 + bZ_2 + cZ_3) = a$이므로

$$a = \rho_{13}.$$

Y_2**와의 상관계수.** $\mathrm{Cov}(\rho_{12}Z_1 + sZ_2,\ aZ_1 + bZ_2 + cZ_3) = \rho_{12}a + sb$이므로 $\rho_{12}\rho_{13} + sb = \rho_{23}$이고, $s \neq 0$이므로

$$b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}.$$

분산이 1. $a^2 + b^2 + c^2 = 1$이고, $c \geq 0$이므로 음이 아닌 근을 택한다.

$$\begin{aligned} c^2 &= 1 - \rho_{13}^2 - \frac{(\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2} \\ &= \frac{(1-\rho_{12}^2)(1-\rho_{13}^2) - (\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2} \\ &= \frac{1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23}}{1-\rho_{12}^2}, \end{aligned}$$

여기서 마지막 단계에서 $\rho_{12}^2\rho_{13}^2$ 항들이 소거된다. 따라서

$$c = \sqrt{\frac{1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23}}{1-\rho_{12}^2}}.$$

생각해 보기 근호 안의 식이 음수이면 어떻게 될까?

그러면 실수 c가 존재하지 않으며, 이 세 상관계수를 갖는 확률변수는 아예 존재하지 않는다. 즉 수 $\rho_{12}, \rho_{13}, \rho_{23}$이 서로 모순된다. 예를 들어 $\rho_{12} = \rho_{13} = 0.9$이고 $\rho_{23} = -0.9$인 것은 불가능하다. Y_2와 Y_3이 둘 다 Y_1에 가깝다면 이 둘이 서로 강하게 반대 방향일 수는 없기 때문이다. 문제는 주어진 상관계수들이 실현 가능하다고 암묵적으로 가정하고 있으며, 이것이 바로 이 분자가 음이 아니라는 조건이다.

마지막 부분: X_i

$i = 1, 2, 3$에 대하여

$$X_i = \mu_i + \sigma_i Y_i,$$

로 놓고, $\sigma_i > 0$으로 한다(σ_i^2만 주어져 있으므로 σ_i를 양수로 택할 수 있다). 그러면 $\mathrm{E}(X_i) = \mu_i + \sigma_i \cdot 0 = \mu_i$이고 $\mathrm{Var}(X_i) = \sigma_i^2 \mathrm{Var}(Y_i) = \sigma_i^2$이다. 상수를 더해도 공분산은 변하지 않으므로

$$\mathrm{Corr}(X_i, X_j) = \frac{\sigma_i \sigma_j \,\mathrm{Cov}(Y_i, Y_j)}{\sqrt{\sigma_i^2 \sigma_j^2}} = \frac{\sigma_i \sigma_j}{|\sigma_i \sigma_j|}\,\rho_{ij} = \rho_{ij},$$

이며, 여기서 $\sigma_i \sigma_j > 0$을 사용하였다. 따라서 이 X_i들은 요구된 성질을 모두 갖는다.

검산

- $\rho_{13} = \rho_{23} = 0$이면 $a = b = 0, c = 1$이므로 $Y_3 = Z_3$이 되어, 마땅히 그래야 하듯이 다른 변수들과 독립이다. $\rho_{12} = 0$이면 $Y_2 = Z_2$이고 $a = \rho_{13}, b = \rho_{23}, c = \sqrt{1 - \rho_{13}^2 - \rho_{23}^2}$이 되어 자명한 선택과 일치한다.
- 수치적으로, $\rho_{12} = 0.5, \rho_{13} = -0.3, \rho_{23} = 0.2$이면 $a = -0.3, b \approx 0.4041, c \approx 0.8641$이고 $a^2 + b^2 + c^2 = 1$이다. 2×10^5개의 세 쌍을 시뮬레이션한 결과(Z_k로 정규분포가 아닌 균등분포를 사용), X_i들의 표본 상관계수는 0.500, −0.302, 0.200이었고 평균과 표준편차도 올바르게 나왔다.

흔한 실수

- $\mathrm{Var}(Y_2) = 1$이므로 상관계수가 곧 공분산이라는 사실을 잊고, 엉뚱한 값으로 나누는 것.
- $X_i = \mu_i + \sigma_i^2 Y_i$로 쓰는 것: 분산은 곱하는 수의 **제곱**만큼 커진다.
- 어떤 변수에 대해 $\sigma_i < 0$으로 택하는 것: 그러면 그 변수의 상관계수 부호가 뒤집힌다.

STEP 너머

세 조건은 계수 행렬 A(행이 $(1,0,0)$, $(\rho_{12},s,0)$, (a,b,c))가 $AA^{\mathsf{T}}=R$을 만족한다는 뜻이며, 여기서 R은 상관행렬(correlation matrix)이다. 이것이 R의 **숄레스키 분해**(*Cholesky decomposition*)이며, 이 분해는 R이 양의 준정부호(positive semi-definite)일 때 정확히 존재한다. c^2의 분자는 $\det R$이다. 이것은 통계학자들이 실제로 상관된 변수들을 시뮬레이션하는 방법이며, Z_k로 정규분포를 사용하면 다변량 정규분포(multivariate normal distribution)가 만들어진다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 공분산, 상관계수, 다변량 정규분포를 다루고, Strang의 *Introduction to Linear Algebra*(선형대수학 입문)는 양의 정부호 행렬과 숄레스키 분해를 설명한다.

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.