解答例

STEP III 2010、第 13 問

確率・統計 ・ 線形結合の相関

STEP III 2010、問 13

先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：

step.maths.org/questions/10-s3-q13

この問題は、*Beyond STEP* の第 13.7 節「**相関をつくり出す**」で詳しく解説されている：
https://beyondstep.keithmcnulty.org/ja/ch13/correlation.html

鍵となる考え方

ここに現れるものはすべて、平均 0、分散 1 の独立な確率変数の線形結合（linear combination）である。したがって、期待値・分散・共分散はすべて係数から直ちに読み取れる。すなわち $U = \sum u_k Z_k$、$V = \sum v_k Z_k$ ならば $\operatorname{Cov}(U, V) = \sum u_k v_k$ である。この問題が実際に行っているのは、指定された任意の相関をもつ三つの標準化された確率変数を一つずつ構成することであり、これは姿を変えたグラム・シュミットの直交化（Gram–Schmidt process）にほかならない。新しい条件が一つ加わるごとに、新しい係数が一つ決まる。

解答

以下では一貫して、共分散の双線形性（bilinearity）、$\operatorname{Var}(Z_k) = 1$、および $j \neq k$ に対する $\operatorname{Cov}(Z_j, Z_k) = 0$（$Z_k$ が独立であることから成り立つ）を用いる。特に、定数 u_k、v_k に対して

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

Z_1 と Z_2 の相関

独立性より $\mathrm{Cov}(Z_1, Z_2) = \mathrm{E}(Z_1Z_2) - \mathrm{E}(Z_1)\mathrm{E}(Z_2) = 0$ であるから、$\mathrm{Corr}(Z_1, Z_2) = 0$ である。

確率変数 Y_2

$s = \sqrt{1-\rho_{12}^2}$ とおく。$-1 < \rho_{12} < 1$ であるから、これは正の実数である。このとき $Y_2 = \rho_{12}Z_1 + sZ_2$ であり、期待値の線形性より

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

上で述べた共分散の規則により、

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

$\mathrm{Var}(Y_1) = \mathrm{Var}(Y_2) = 1$ であるから、定義式の分母は 1 となり、

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

確率変数 Y_3

条件 $\mathrm{E}(Y_3) = 0$ はどのような a, b, c に対しても自動的に成り立つので、何の情報も与えない。残りの三つの条件はそれぞれ一つずつ方程式を与える。Y_1, Y_2, Y_3 の分散はすべて 1 であるから、各相関は対応する共分散に等しい。

Y_1 との相関。 $\mathrm{Cov}(Z_1,\, aZ_1 + bZ_2 + cZ_3) = a$ であるから、

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

Y_2 との相関。 $\mathrm{Cov}(\rho_{12}Z_1 + sZ_2,\, aZ_1 + bZ_2 + cZ_3) = \rho_{12}a + sb$ であるから、$\rho_{12}\rho_{13} + sb = \rho_{23}$ となり、$s \neq 0$ より

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

分散が 1 であること。 $a^2 + b^2 + c^2 = 1$ であり、$c \geq 0$ であるから非負の平方根をとる。

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

ここで最後の段階では $\rho_{12}^2\rho_{13}^2$ の項が打ち消し合う。よって

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

考えてみよう 平方根の中の式が負になったらどうなるか？

その場合、実数 c は存在せず、これら三つの相関をもつ確率変数はそもそも存在しない。すなわち、数 $\rho_{12},\rho_{13},\rho_{23}$ は互いに整合していない。たとえば $\rho_{12}=\rho_{13}=0.9$ かつ $\rho_{23}=-0.9$ は不可能である。Y_2 と Y_3 がともに Y_1 に近いならば、両者が互いに強く逆向きであることはありえないからである。この問題は、与えられた相関が実現可能であることを暗黙のうちに仮定しており、それはちょうどこの分子が非負であるという条件にあたる。

最後の小問：X_i

$i=1,2,3$ に対して

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

とおく。ただし $\sigma_i>0$ とする（指定されているのは σ_i^2 だけなので、σ_i は正にとってよい）。このとき $\mathrm{E}(X_i)=\mu_i+\sigma_i\cdot 0=\mu_i$ かつ $\mathrm{Var}(X_i)=\sigma_i^2\mathrm{Var}(Y_i)=\sigma_i^2$ である。定数を加えても共分散は変わらないので、

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

となる。ここで $\sigma_i\sigma_j>0$ を用いた。したがって、これらの X_i は要求された性質をすべてもつ。

検算

- $\rho_{13}=\rho_{23}=0$ ならば $a=b=0$、$c=1$ となるので $Y_3=Z_3$ であり、当然そうあるべきとおり他の変数と独立になる。$\rho_{12}=0$ ならば $Y_2=Z_2$ であり、$a=\rho_{13}$、$b=\rho_{23}$、$c=\sqrt{1-\rho_{13}^2-\rho_{23}^2}$ となって、自然な選び方に一致する。
- 数値的には、$\rho_{12}=0.5$、$\rho_{13}=-0.3$、$\rho_{23}=0.2$ のとき $a=-0.3$、$b\approx 0.4041$、$c\approx 0.8641$ となり、$a^2+b^2+c^2=1$ が成り立つ。2×10^5 組の三つ組のシミュレーション（Z_k には正規分布ではなく一様分布を用いた）では、X_i の標本相関は 0.500、-0.302、0.200 となり、平均と標準偏差も正しい値であった。

よくある間違い

- $\mathrm{Var}(Y_2) = 1$ であるから相関は単に共分散に等しい、ということを忘れ、誤ったもので割ってしまう。
- $X_i = \mu_i + \sigma_i^2 Y_i$ と書いてしまう。分散は掛ける数の **2 乗**倍になる。
- ある変数について $\sigma_i < 0$ ととってしまい、その変数の相関の符号が反転する。

STEP のその先へ

三つの条件は、係数の行列 A（行が $(1,0,0)$、$(\rho_{12}, s, 0)$、(a,b,c)）が $AA^{\mathsf{T}} = R$ を満たすことを意味する。ここで R は相関行列（correlation matrix）である。これは R の**コレスキー分解**（Cholesky decomposition）であり、R が半正定値（positive semi-definite）であるときに限り存在する。c^2 の分子は $\det R$ である。これは統計学者が実際に相関のある確率変数をシミュレーションする方法であり、Z_k を正規分布にとれば多変量正規分布（multivariate normal distribution）が得られる。Blitzstein and Hwang, *Introduction to Probability*（確率論入門）は共分散、相関、多変量正規分布を扱っている。Strang, *Introduction to Linear Algebra*（線形代数入門）は正定値行列とコレスキー分解を解説している。

Beyond STEP 解答集：https://beyondstep.keithmcnulty.org/ja/。解答本文は CC0 1.0 である。問題そのものは © Cambridge University Press & Assessment に帰属し、ここには掲載していない。英語版から機械翻訳したものであり、まだ校閲を経ていない。