详细解答

STEP III 2002，第 14 题

概率与统计 · 方差与协方差

STEP III 2002，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/02-s3-q14

关键思路

这条方差恒等式其实就是方差（variance）的定义，再利用期望值的线性性质（linearity of expectation）展开而已。国际象棋那部分比表面看来更巧妙：我们从头到尾都没有直接计算 $\mathrm{E}(WL)$。相反，我们由二项分布（binomial distribution）得知 $W+L$ 的方差，同样由二项分布得知 $\mathrm{Var}(W)$ 和 $\mathrm{Var}(L)$，而刚证明的恒等式便把协方差（covariance）白白送到我们手上。

解答

方差恒等式

记 $\mu_X = \mathrm{E}(X)$ 及 $\mu_Y = \mathrm{E}(Y)$。我们采用以下定义

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu_X)^2], \qquad \mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)],$$

以及期望值的线性性质 $\mathrm{E}(aU+bV) = a\,\mathrm{E}(U) + b\,\mathrm{E}(V)$；对于离散随机变量（discrete random variable），只要把和式 $\sum (au+bv)\,\mathrm{P}(U=u, V=v)$ 拆成两个和式，并在每个和式中对另一个变量求和，便可得出这个性质。

根据线性性质，$\mathrm{E}(X+Y) = \mu_X + \mu_Y$，所以

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}[((X-\mu_X)+(Y-\mu_Y))^2] \\
&= \mathrm{E}[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] \\
&= \mathrm{E}[(X-\mu_X)^2] + \mathrm{E}[(Y-\mu_Y)^2] + 2\,\mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y),
\end{aligned}$$

其中第三行再次用到线性性质。这就是所需的结果。留意我们并不需要独立性（independence）：这条恒等式对任何 X 和 Y 都成立。

第 (i) 部分

每一局棋不是胜、就是负、或者和，所以无论结果如何，都有 $W + L + D = m$。一个永远取同一个值的随机变量等于它的平均值，所以它的方差是 $\mathrm{E}[(m-m)^2] = 0$。

第 (ii) 部分

如果一局棋分出了胜负（即胜或负），我们称它为*分胜负*的一局。每一局分胜负的概率是 $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$，各局互相独立，而且局数固定为 m。因此 $W + L$，即分胜负的局数，就是在 m 次成功概率相同的独立试验中成功的次数：

$$W + L \sim \mathrm{B}(m, \tfrac{2}{3}).$$

相关系数

同样道理，W 是 m 次独立试验中胜出的次数，每次胜出的概率是 $\frac{1}{3}$，所以 $W \sim \mathrm{B}(m, \frac{1}{3})$，同理 $L \sim \mathrm{B}(m, \frac{1}{3})$。利用二项分布的公式 $\mathrm{Var} = np(1-p)$，

$$\mathrm{Var}(W) = \mathrm{Var}(L) = m \cdot \tfrac{1}{3} \cdot \tfrac{2}{3} = \frac{2m}{9}, \qquad \mathrm{Var}(W+L) = m \cdot \tfrac{2}{3} \cdot \tfrac{1}{3} = \frac{2m}{9}.$$

想一想 怎样可以不计算 $\mathrm{E}(WL)$ 而求出 $\mathrm{Cov}(W, L)$?

把恒等式倒过来用。取 $X = W$ 及 $Y = L$，

$$\frac{2m}{9} = \frac{2m}{9} + \frac{2m}{9} + 2\,\mathrm{Cov}(W, L), \qquad \text{所以} \qquad \mathrm{Cov}(W, L) = -\frac{m}{9}.$$

因此，当 $m \geq 1$ 时（这样方差才不是零），

$$\frac{\mathrm{Cov}(W, L)}{\sqrt{\mathrm{Var}(W)\,\mathrm{Var}(L)}} = \frac{-m/9}{2m/9} = -\frac{1}{2}.$$

答案与 m 无关。它是负数，这亦合乎情理：局数既然固定，胜的局数越多，留给负局的空间便越少。

第 (i) 部分提供了另一条途径。把恒等式应用两次，可知 $\mathrm{Var}(W + L + D)$ 等于三个方差之和，再加上三个两两协方差之和的两倍。根据对称性，三个协方差全部相等，设为 c，于是 $0 = 3 \cdot \frac{2m}{9} + 6c$，再次得出 $c = -\frac{m}{9}$。

验算

- 取 $m = 1$。这时 W 和 L 都是指示变量（indicator），$\mathrm{E}(W) = \mathrm{E}(L) = \frac{1}{3}$，而且恒有 $WL = 0$，所以 $\mathrm{Cov}(W, L) = 0 - \frac{1}{9} = -\frac{1}{9}$，与 $-m/9$ 吻合。
- 两条途径（经第 (ii) 部分和经第 (i) 部分）的结果一致。
- 对 $m = 1, \ldots, 7$ 逐一穷举全部 3^m 个结果，每次都得出 $\mathrm{Var}(W) = 2m/9$、$\mathrm{Cov}(W, L) = -m/9$，而相关系数（correlation）都恰好是 $-\frac{1}{2}$。

常见错误

- 因为各局棋互相独立，便以为 W 和 L 也互相独立。其实不然：它们与 D 合起来的总数是固定的。
- 取 $\mathrm{Var}(W + L) = \mathrm{Var}(W) + \mathrm{Var}(L) = 4m/9$，这样做漏掉了协方差一项，而且与二项分布得出的答案 $2m/9$ 矛盾。

STEP 之外

(W, L, D) 服从多项分布（multinomial distribution），即把二项分布自然地推广至每次试验有多于两种结果的情况。一般而言，若各种结果的概率为 $p_1, \ldots, p_k$，则计数 N_i 与 N_j 的协方差为 $-mp_ip_j$，相关系数为 $-\sqrt{p_ip_j/\big((1-p_i)(1-p_j)\big)}$，在本题中就是 $-\frac{1}{2}$。协方差矩阵（covariance matrix）及协方差的双线性（bilinearity），可参阅 Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门），以及 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论导论）。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。