Topic. Sample correlation coefficient

Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots$ are iid vectors with E $X_i^4 < \infty$ and E $Y_i^4 < \infty$. For the sake of simplicity, we will assume without loss of generality that E X_i = E $Y_i = 0$ (alternatively, we could base all of the following derivations on the centered versions of the random variables).

We wish to find the asymptotic distribution of the sample correlation $r = s_{xy}/(s_x s_y)$, where if we let

$$\begin{pmatrix} m_x \\ m_y \\ m_{xx} \\ m_{yy} \\ m_{xy} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i^2 \\ \sum_{i=1}^n Y_i^2 \\ \sum_{i=1}^n X_i Y_i \end{pmatrix}, \tag{35}$$

then

$$s_x^2 = m_{xx} - m_x^2, s_y^2 = m_{yy} - m_y^2, \text{ and } s_{xy} = m_{xy} - m_x m_y. \tag{36}$$

Notice that we have suppressed the n in the notation above in order to keep things slightly simpler. According to the central limit theorem,

$$\sqrt{n} \left\{ \begin{pmatrix} m_x \\ m_y \\ m_{xx} \\ m_{yy} \\ m_{xy} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right\} \xrightarrow{\mathcal{L}} N_5 \left\{ \underline{0}, \begin{pmatrix} \text{Cov}\,(X_1, X_1) & \cdots & \text{Cov}\,(X_1, X_1 Y_1) \\ \text{Cov}\,(Y_1, X_1) & \cdots & \text{Cov}\,(Y_1, X_1 Y_1) \\ \vdots & \ddots & \vdots \\ \text{Cov}\,(X_1 Y_1, X_1) & \cdots & \text{Cov}\,(X_1 Y_1, X_1 Y_1) \end{pmatrix} \right\}. \tag{37}$$

Let Σ denote the covariance matrix in expression (37). Define a function $g : R^5 \to R^3$ such that g applied to the vector of moments in equation (35) yields the vector (s_x^2, s_y^2, s_{xy}) as defined in expression (36). Then

$$\dot{g} \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} -2a & 0 & 1 & 0 & 0 \\ 0 & -2b & 0 & 1 & 0 \\ -b & -a & 0 & 0 & 1 \end{pmatrix}.$$

Therefore, if we let

$$\Sigma^* = \dot{g} \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \Sigma \dot{g} \begin{pmatrix} 0 \\ 0 \\ \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix}^t = \begin{pmatrix} \text{Cov}\,(X_1^2, X_1^2) & \text{Cov}\,(X_1^2, Y_1^2) & \text{Cov}\,(X_1^2, X_1 Y_1) \\ \text{Cov}\,(Y_1^2, X_1^2) & \text{Cov}\,(Y_1^2, Y_1^2) & \text{Cov}\,(Y_1^2, X_1 Y_1) \\ \text{Cov}\,(X_1 Y_1, X_1^2) & \text{Cov}\,(X_1 Y_1, Y_1^2) & \text{Cov}\,(X_1 Y_1, X_1 Y_1) \end{pmatrix},$$

then by the delta method,

$$\sqrt{n} \left\{ \begin{pmatrix} s_x^2 \\ s_y^2 \\ s_{xy} \end{pmatrix} - \begin{pmatrix} \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} \right\} \xrightarrow{\mathcal{L}} N_3(\underline{0}, \Sigma^*).$$

Finally, define the function $h(a,b,c) = c/\sqrt{ab}$, so that we have $h(s_x^2, s_y^2, s_{xy}) = r$. Then $\dot{h}(a,b,c) = \frac{1}{2}(-c/\sqrt{a^3 b}, -c/\sqrt{ab^3}, 2/\sqrt{ab})$, so that

$$\dot{h}\begin{pmatrix} \sigma_x^2 \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} = \left(\frac{-\sigma_{xy}}{2\sigma_x^3\sigma_y}, \frac{-\sigma_{xy}}{2\sigma_x\sigma_y^3}, \frac{1}{\sigma_x\sigma_y} \right) = \left(\frac{-\rho}{2\sigma_x^2}, \frac{-\rho}{2\sigma_y^2}, \frac{1}{\sigma_x\sigma_y} \right). \tag{38}$$

Therefore, if A denotes the 1×3 matrix in equation (38), using the delta method once again yields

$$\sqrt{n}(r - \rho) \xrightarrow{\mathcal{L}} N(0, A\Sigma^* A^t).$$

Consider the special case of bivariate normal (X_i, Y_i). In this case, we may derive

$$\Sigma^* = \begin{pmatrix} 2\sigma_x^4 & 2\rho^2\sigma_x^2\sigma_y^2 & 2\rho\sigma_x^3\sigma_y \\ 2\rho^2\sigma_x^2\sigma_y^2 & 2\sigma_y^2 & 2\rho\sigma_x\sigma_y^3 \\ 2\rho\sigma_x^3\sigma_y & 2\rho\sigma_x\sigma_y^3 & (1+\rho^2)\sigma_x^2\sigma_y^2 \end{pmatrix}. \tag{39}$$

In this case, $A\Sigma^* A^t = (1-\rho^2)^2$, which implies that

$$\sqrt{n}(r - \rho) \xrightarrow{\mathcal{L}} N\{0, (1-\rho^2)^2\}. \tag{40}$$

In the normal case, we may derive a variance-stabilizing transformation. According to equation (40), we should find a function $f(x)$ satisfying $f'(x) = (1-x^2)^{-1}$. Since

$$\frac{1}{1-x^2} = \frac{1}{2(1-x)} + \frac{1}{2(1+x)},$$

which is easy to integrate, we obtain

$$f(x) = \frac{1}{2}\log\frac{1+x}{1-x}.$$

This is called Fisher's transformation; we conclude that

$$\sqrt{n}\left(\frac{1}{2}\log\frac{1+r}{1-r} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}\right) \xrightarrow{\mathcal{L}} N(0,1).$$