19. Sample correlation coefficient

Lehmann §5.4; Ferguson §8

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

\[
\begin{pmatrix}
    m_x \\
    m_y \\
    m_{xx} \\
    m_{yy} \\
\end{pmatrix} = \frac{1}{n} \begin{pmatrix}
    \sum_{i=1}^n X_i \\
    \sum_{i=1}^n Y_i \\
    \sum_{i=1}^n X_i^2 \\
    \sum_{i=1}^n Y_i^2 \\
\end{pmatrix},
\]

then

\[s_x^2 = m_{xx} - m_x^2, s_y^2 = m_{yy} - m_y^2, \quad \text{and} \quad s_{xy} = m_{xy} - m_x m_y.\]  

(38)

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

\[
\sqrt{n} \begin{pmatrix}
    m_x \\
    m_y \\
    m_{xx} \\
    m_{yy} \\
\end{pmatrix} - \begin{pmatrix}
    0 \\
    0 \\
    \sigma_x^2 \\
    \sigma_y^2 \\
\end{pmatrix} \sim N_5 \begin{pmatrix}
    \text{Cov}(X_1, X_1) & \cdots & \text{Cov}(X_1, X_1 Y_1) \\
    \text{Cov}(Y_1, X_1) & \cdots & \text{Cov}(Y_1, X_1 Y_1) \\
    \vdots & \ddots & \vdots \\
    \text{Cov}(X_1 Y_1, X_1) & \cdots & \text{Cov}(X_1 Y_1, X_1 Y_1) \\
\end{pmatrix}.
\]

(39)

Let \(\Sigma\) denote the covariance matrix in expression (39). Define a function \(g : \mathbb{R}^5 \to \mathbb{R}^3\) such that \(g\) applied to the vector of moments in equation (37) yields the vector \((s_x^2, s_y^2, s_{xy})\) as defined in expression (38). Then

\[
\begin{pmatrix}
    a \\
    b \\
    c \\
    d \\
    e \\
\end{pmatrix} = \begin{pmatrix}
    -2a \\
    0 \\
    -2b \\
    0 \\
    -a \\
\end{pmatrix}.
\]

Therefore, if we let

\[
\Sigma^* = \hat{g} \begin{pmatrix}
    0 \\
    0 \\
    \sigma_x^2 \\
    \sigma_y^2 \\
    \sigma_{xy} \\
\end{pmatrix} \Sigma \hat{g} \begin{pmatrix}
    0 \\
    0 \\
    \sigma_x^2 \\
    \sigma_y^2 \\
    \sigma_{xy} \\
\end{pmatrix}^\top = \begin{pmatrix}
    \text{Cov}(X_1^2, X_1^2) & \text{Cov}(X_1^2, X_1 Y_1) \\
    \text{Cov}(Y_1^2, X_1^2) & \text{Cov}(Y_1^2, Y_1^2) \\
    \text{Cov}(X_1 Y_1, X_1^2) & \text{Cov}(X_1 Y_1, X_1 Y_1) \\
\end{pmatrix},
\]

then by the delta method,

\[
\sqrt{n} \begin{pmatrix}
    s_x^2 \\
    s_y^2 \\
    s_{xy} \\
\end{pmatrix} - \begin{pmatrix}
    \sigma_x^2 \\
    \sigma_y^2 \\
    \sigma_{xy} \\
\end{pmatrix} \sim N_3(0, \Sigma^*).
\]

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

\[
\hat{h} \begin{pmatrix}
    \sigma_x^2 \\
    \sigma_y^2 \\
    \sigma_{xy} \\
\end{pmatrix} = \begin{pmatrix}
    -\sigma_{xy} \\
    -\sigma_{xy} \\
    1 \\
\end{pmatrix} \begin{pmatrix}
    \frac{1}{2} \sigma_x^2 \\
    \frac{1}{2} \sigma_y^2 \\
    \sigma_{xy} \\
\end{pmatrix} = \begin{pmatrix}
    -\rho \\
    -\rho \\
    1 \\
\end{pmatrix} 
\]

(40)

52
Therefore, if $A$ denotes the $1 \times 3$ matrix in equation (40), using the delta method once again yields

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

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

$$\Sigma^* = \begin{pmatrix}
2\sigma_x^4 & 2\rho\sigma_x^2\sigma_y^2 & 2\rho\sigma_x^2\sigma_y^2 \\
2\rho\sigma_x^2\sigma_y^2 & 2\sigma_y^4 & 2\rho\sigma_x\sigma_y^3 \\
2\rho\sigma_x^2\sigma_y^2 & 2\rho\sigma_x\sigma_y^3 & (1 + \rho^2)\sigma_x^2\sigma_y^2
\end{pmatrix}. \quad (41)$$

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

$$\sqrt{n}(r - \rho) \xrightarrow{d} N(0, (1 - \rho^2)^2). \quad (42)$$

In the normal case, we may derive a variance-stabilizing transformation. According to equation (42), 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

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

### Problems

**Problem 19.1** Verify expressions (41) and (42).

**Problem 19.2** Assume $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid from some bivariate normal distribution. Let $\rho$ denote the population correlation coefficient and $r$ the sample correlation coefficient.

(a) Describe a test of $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ based on the fact that

$$\sqrt{n}[f(r) - f(\rho)] \xrightarrow{d} N(0, 1),$$

where $f(x)$ is Fisher’s transformation $f(x) = (1/2) \log[(1 + x)/(1 - x)]$. Use $\alpha = .05$.

(b) Based on 5000 repetitions each, estimate the actual level for this test in the case when $E(X_i) = E(Y_i) = 0$, $\text{Var}(X_i) = \text{Var}(Y_i) = 1$, and $n \in \{3, 5, 10, 20\}$.

**Problem 19.3** Suppose that $X$ and $Y$ are jointly distributed such that $X$ and $Y$ are Bernoulli (1/2) random variables with $P(XY = 1) = \theta$ for $\theta \in (0, 1/2)$. Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be iid with $(X_i, Y_i)$ distributed as $(X, Y)$.

(a) Find the asymptotic distribution of $\sqrt{n} \left[ (\bar{X}_n, \bar{Y}_n) - (1/2, 1/2) \right]$.

(b) If $r_n$ is the sample correlation coefficient for a sample of size $n$, find the asymptotic distribution of $\sqrt{n}(r_n - \rho)$. 

53
(c) Find a variance stabilizing transformation for $r_n$.

(d) Based on your answer to part (c), construct a 95% confidence interval for $\theta$.

(e) For each combination of $n \in \{5, 20\}$ and $\theta \in \{.05, .25, .45\}$, estimate the true coverage probability of the confidence interval in part (d) by simulating 5000 samples and the corresponding confidence intervals.

**Hint:** To generate a sample of $(X, Y)$, first simulate the $X$'s from their marginal distribution, then simulate the $Y$'s according to the conditional distribution of $Y$ given $X$. To obtain this conditional distribution, simply find $P(Y = 1 \mid X = 1)$ and $P(Y = 1 \mid X = 0)$ using the definition of conditional probability.