Covariance and Correlation

November, 2009

Here, we shall assume that the random variables under consideration have positive and finite variance.

One simple way to assess the relationship between two random variables X and Y is to compute their **covariance**.

$$\mathrm{Cov}(X,Y) = E[(X-\mu_x)(Y-\mu_y)].$$

Exercise 1. $\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X,Y)$. *and*

$$\mathrm{Var}(aX+cY) = a^2\mathrm{Var}(X) + 2ac\mathrm{Cov}(X,Y) + c^2\mathrm{Var}(Y). \tag{1}$$

As with the variance, we have an alternative definition of covariance.

$$\mathrm{Cov}(X,Y) = EXY - \mu_Y EX - \mu_X EY + \mu_X\mu_Y = EXY - \mu_X mu_Y.$$

Example 2. *For the joint density example,*

$$\begin{aligned} EXY &= \frac{4}{5}\int_0^1\int_0^1 xy(x+y+xy)\,dy\,dx = \frac{4}{5}\int_0^1\int_0^1 (x^2y+xy^2+x^2y^2)\,dy\,dx \\ &= \frac{4}{5}\int_0^1 \left(\frac{1}{2}x^2y^2 + \frac{1}{3}xy^3 + \frac{1}{3}x^2y^3\right)\Big|_0^1 dx = \frac{4}{5}\int_0^1\left(\frac{5}{6}x^2+\frac{1}{3}x\right)dx \\ &= \frac{4}{5}\left(\frac{5}{18}x^3+\frac{1}{6}x^2\right)\Big|_0^1 = \frac{4}{5}\left(\frac{5}{18}+\frac{1}{6}\right) = \frac{16}{45} \end{aligned}$$

$$EX = EY = \frac{2}{5}\int_0^1 x(3x+1)\,dx = \frac{2}{5}\left(x^3+\frac{1}{2}x^2\right)\Big|_0^1 = \frac{2}{5}\cdot\frac{3}{2} = \frac{3}{5}.$$

$$\mathrm{Cov}(X,Y) = \frac{16}{45} - \left(\frac{3}{5}\right)^2 = \frac{80-81}{225} = -\frac{1}{225}.$$

The **correlation** is the covariance of the standardized version of the random variables.

$$\rho_{X,Y} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

In the example,

$$\sigma_X^2 = \frac{2}{5}\int_0^1 x^2(3x+1)\,dx - \left(\frac{3}{5}\right)^2 = \frac{2}{5}\cdot\frac{13}{12} - \frac{9}{25} = \frac{11}{150}.$$

and

$$\rho_{X,Y} = \frac{-1/225}{11/150} = -\frac{2}{33} = -0.06.$$

We can write equation (1) with $a = 1$ as

$$\sigma^2_{X+cY} = \sigma^2_X + 2\rho_{X,Y}\sigma_X\sigma_Y c + \sigma^2_Y c^2.$$

This must be nonnegative for all values of c. Thus, by considering the quadratic formula, we have that the discriminate

$$0 \geq (2\rho_{X,Y}\sigma_X\sigma_Y)^2 - 4\sigma^2_X\sigma^2_Y = (\rho^2_{X,Y} - 1)4\sigma^2_X\sigma^2_Y \quad \text{or} \quad \rho^2_{X,Y} \leq 1.$$

Consequently,

$$-1 \leq \rho_{X,Y} \leq 1.$$

When we have $|\rho_{X,Y}| = 1$, we also have for some value of c that

$$\sigma^2_{X+cY} = 0.$$

In this case, $X + cY$ is a constant random variable and X and Y are linearly related. In this case, the sign of $\rho_{X,Y}$ depends on the sign of the linear relationship.

Exercise 3. $\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j)$.

Example 4 (variance of a hypergeometric)**.** *Consider an urn with B blue balls and G green balls. Remove K and let the random variable X denote the number of blue balls. Let*

$$X_i = \begin{cases} 0 & \text{if the } i\text{-th ball is green,} \\ 1 & \text{if the } i\text{-th ball is blue.} \end{cases}$$

Then, $X = X_1 + X_2 + \cdots + X_K$. First, note that X_i is a Bernoulli random variable. $EX_i = B/(B+G)$ and $\text{Var}(X_i) = BG/(B+G)^2$. *Next, for the $K(K-1)$ terms with $i \neq j$,*

$$E[X_iX_j] = P\{X_i = 1, X_j = 1\} = P\{X_i = 1 | X_j = 1\}P\{X_j = 1\} = \frac{B-1}{B+G-1} \cdot \frac{B}{B+G}.$$

Thus,

$$\begin{aligned} \text{Cov}(X_i, X_j) &= \frac{B(B-1)}{(B+G)(B+G-1)} - \left(\frac{B}{B+G}\right)^2 = \frac{B}{B+G}\left(\frac{B-1}{B+G-1} - \frac{B}{B+G}\right) \\ &= \frac{B}{B+G}\left(\frac{-G}{(B+G)(B+G-1)}\right) = \frac{-BG}{(B+G)^2(B+G-1)} \end{aligned}$$

and using the formula in the previous exercise with the $a_i = 1$,

$$\text{Var}(X) = K\frac{BG}{(B+G)^2} + K(K-1)\left(\frac{-BG}{(B+G)^2(B+G-1)}\right) = K\frac{BG}{(B+G)^2}\left(1 - \frac{K-1}{B+G-1}\right).$$

To simplify the appearance of this expression, let $N = K+G$ be the total number of balls and $p = B/(B+G)$ be the proportion of the total number of balls that are blue. Then,

$$\text{Var}(X) = Kp(1-p)\frac{N-K}{N-1}.$$

Note that if $K << N$, then the variance is essentially the same as that of the corresponding binomial random variable. At the other extreme, if $K = N$, then all the balls have been removed from the urn and $Var(X) = 0$.