KTH Matematik

För ett datamaterial av n talpar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ definieras kovariansen av

$$\text{Kov} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

och korrelationskoefficienten av

$$\rho = \frac{\text{Kov}}{s_x s_y}$$

där $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ är standardavvikelsen för x_i-värdena och motsvarande för s_y. Om man utnyttjar att $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = n\overline{x} - n\overline{x} = 0$ erhåller man alternativa sätt att beräkna Kov:

$$\begin{aligned}\text{Kov} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})y_i - \overline{y}\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})y_i = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}\Big).\end{aligned}$$

För att visa att korrelationskoefficienten ligger mellan -1 och $+1$ kan man utnyttja Schwartz olikhet:

Sats Låt $a_1, a_2, \ldots, a_n$ och $b_1, b_2, \ldots, b_n$ var godtyckliga rellea tal. Då är

$$\Big(\sum_{i=1}^{n} a_i b_i\Big)^2 \le \sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2.$$

Bevis: Låt k vara ett godtyckligt tal. Vi har att

$$0 \le \sum_{i=1}^{n}(a_i - k b_i)^2 = \sum_{i=1}^{n}(a_i^2 - 2k a_i b_i + k^2 b_i^2) = \sum_{i=1}^{n} a_i^2 - 2k\sum_{i=1}^{n} a_i b_i + k^2\sum_{i=1}^{n} b_i^2.$$

Genom att derivera högerledet med avseende på k erhåller man lätt att minimum fås för $k = \sum_{i=1}^{n} a_i b_i / \sum_{i=1}^{n} b_i^2$. Insätter man detta värde i olikheten erhålls

$$0 \le \sum_{i=1}^{n} a_i^2 - 2\frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} b_i^2}\sum_{i=1}^{n} a_i b_i + \left(\frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} b_i^2}\right)^2 \sum_{i=1}^{n} b_i^2 = \sum_{i=1}^{n} a_i^2 - \frac{\left(\sum_{i=1}^{n} a_i b_i\right)^2}{\sum_{i=1}^{n} b_i^2}$$

vilket ger satsen.

Man ser också direkt från beviset att likhet fås om och endast om $a_i = k b_i$, dvs a_i:na proportionella mot b_i:värdena.

Om man i satsen ovan sedan sätter in $a_i = x_i - \overline{x}$ och $b_i = y_i - \overline{y}$ erhålls

$$\rho^2 = \frac{\text{Kov}^2}{s_x^2 s_y^2} = \frac{\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\left(\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \le 1$$

Likhet fås om och endast om $x_i - \overline{x} = k(y_i - \overline{y})$ dvs om och endast om data ligger på en rät linje med lutning k.