Kovarianz und Korrelation

Generell setzen wir eine bivariate (d.h. 2-merkmalige) Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$ der Länge n von metrischen Merkmalen voraus. Für die Berechnung der bivariaten Kennwerte legen wir die folgenden Abkürzungen für arithmetische Mittel fest:

Arithmetische Mittel der x-und y-Merkmale: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, und $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

Arithmetische Mittel der quadrierten x-und y-Merkmale: $\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} {x_i}^2$, und $\overline{y^2} = \frac{1}{n} \sum_{i=1}^{n} {y_i}^2$.

Arithmetische Mittel des Produktes der x-und y-Merkmale: $\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$.

Für Varianz und Standardabweichung benutzen wir die bekannten Abkürzungen und Zusammenhänge:

$\tilde{s}_x^2 = \overline{x^2} - \overline{x}^2$, $\tilde{s}_x = \sqrt{\tilde{s}_x^2}$.

Definition

(Empirische) *Kovarianz*: $\tilde{s}_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})$.

(Empirischer) *Korrelationskoeffizient nach Pearson*: $r_{xy} = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y}$, für $\tilde{s}_x \neq 0$ und $\tilde{s}_y \neq 0$.

Wichtige Eigenschaften der Kovarianz und Korrelation

(1) $\tilde{s}_{xy}$ und r_{xy} messen den linearen Zusammenhang der beiden Merkmale, d.h. wie nahe die Datenpunkte einer Geraden mit der Gleichung $y = m \cdot x + q$ mit $m \neq 0$ kommen.

(2) $-1 \leq r_{xy} \leq 1$

(3) $r_{xy} = 0$ bzw. $\tilde{s}_{xy} = 0$ oder nahe bei Null, bedeutet, dass die Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ gleichmässig um den Schwerpunkt $(\overline{x}, \overline{y})$ verteilt sind.

(4) $r_{xy} = 1$ bedeutet, dass ein positiver linearer Zusammenhang zwischen den Merkmalen besteht.

(5) $r_{xy} = -1$ bedeutet, dass ein negativer linearer Zusammenhang zwischen den Merkmalen besteht.

(6) $r_{xy} > 0$: Die Punkte liegen tendenziell um eine Gerade mit positiver Steigung (gleichsinniger linearer Zusammenhang, positive Korrelation).

(7) $r_{xy} < 0$: Die Punkte liegen tendenziell um eine Gerade mit negativer Steigung (gleichsinniger linearer Zusammenhang, negative Korrelation).

(8) r_{xy} ist nicht robust, d.h. bereits ein Ausreisser in den Daten kann den Wert stark beeinflussen.

(9) $\tilde{s}_{xy} = \overline{xy} - \overline{x} \cdot \overline{y}$.

(10) $r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2} \cdot \sqrt{\overline{y^2} - \overline{y}^2}}$.

Rangkorrelation

Sei $(x_1, \ldots, x_n)$ eine Stichprobe eines metrischen Merkmals.

Der *Rang* $rg(x_i)$ des Stichprobenwertes x_i ist definiert als der Index von x_i in der nach der Grösse geordneten Stichprobe, wenn dieser Wert nur einmal auftritt. Tritt der Stichprobenwert x_i mehrmals auf (man sagt dann, dass *Bindungen* in den Daten auftreten), so ist $rg(x_i)$ das arithmetische Mittel der Indizes aller Stichprobenwerte x_i in der geordneten Stichprobe. Für die Stichprobe $(6, 3, 4, 3, 3, 2, 6)$ erhält man die nach der Grösse geordnete Stichprobe $(2, 3, 3, 3, 4, 6, 6)$ und die Ränge $rg(2) = 1$, $rg(3) = \frac{1}{3}(2+3+4) = 3$, $rg(4) = 5$, $rg(6) = \frac{1}{2}(6+7) = 6.5$.

Definition

Wir setzen eine bivariate Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$ der Länge n von metrischen Merkmalen voraus. Dazu bilden wir die zugehörigen Rangfolgen:

Die Folge der Rangpaare: $rg(xy) = ((rg(x_1), rg(y_1)), \ldots, (rg(x_n), rg(y_n))$

Die Folge der Ränge der x-Werte: $rg(x) = (rg(x_1), \ldots, rg(x_n))$

Die Folge der Ränge der y-Werte: $rg(y) = (rg(y_1), \ldots, rg(y_n))$.

Der (empirische) *Korrelationskoeffizient nach Spearman (Rangkorrelationskoeffizient)* ist definiert als der Pearson Korrelationskoeffizient der Rangfolgen: $r_{Sp} = r_{rg(xy)}$.

Wichtige Eigenschaften der Rangkorrelation

(1) $\tilde{s}_{rg(xy)}$ und r_{Sp} messen den monotonen Zusammenhang der beiden Merkmale, d.h. wie nahe die Datenpunkte einer streng monotonen Funktion kommen. In anderen Worten, es wird gemessen wie gut die Rangordnungen in den x und y Werten sich entsprechen.

(2) $-1 \leq r_{Sp} \leq 1$

(3) $r_{Sp} = 0$ bzw. $\tilde{s}_{rg(xy)} = 0$ bedeutet, dass die Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ gleichmässig um den Schwerpunkt der Ränge$(\overline{rg(x)}, \overline{rg(y)}) = (\frac{n+1}{2}, \frac{n+1}{2})$verteilt sind.

(4) $r_{Sp} = 1$ bedeutet, dass ein streng monoton wachsender funktionaler Zusammenhang zwischen den Merkmalen besteht.

(5) $r_{Sp} = -1$ bedeutet, dass ein streng monoton fallender funktionaler Zusammenhang zwischen den Merkmalen besteht.

(6) r_{Sp} ist robust, d.h. Ausreisser in den Datenpunkten beeinflussen den Wert nicht.

(7) $\tilde{s}_{rg(xy)} = \overline{rg(xy)} - \overline{rg(x)} \cdot \overline{rg(y)} = \overline{rg(xy)} - \frac{(n+1)^2}{4}$.

(8) $r_{Sp} = \frac{\overline{rg(xy)} - \overline{rg(x)} \cdot \overline{rg(y)}}{\sqrt{\overline{rg(x)^2} - \overline{rg(x)}^2} \cdot \sqrt{\overline{rg(y)^2} - \overline{rg(y)}^2}} = \frac{\overline{rg(xy)} - \frac{(n+1)^2}{4}}{\sqrt{\overline{rg(x)^2} - \frac{(n+1)^2}{4}} \cdot \sqrt{\overline{rg(y)^2} - \frac{(n+1)^2}{4}}}$.

(9) Falls jeder Stichprobenwert der x und der y-Werte nur einmal vorkommt, in anderen Worten, falls die Ränge in den x Werte und auch in den y-Werte alle verschieden sind (also keine Bindungen auftreten), gibt es eine einfache Berechnungsformel: $r_{Sp} = 1 - \frac{6 \cdot \sum_{i=1}^{n} {d_i}^2}{n \cdot (n^2 - 1)}$ mit $d_i = rg(y_i) - rg(x_i)$.