Der Korrelationskoeffizient nach Pearson

1 Motivation

In der Statistik werden wir uns häufig mit empirisch erfassten Daten beschäftigen. Um diese auszuwerten, ist es oftmals notwendig einen Zusammenhang zwischen den Daten herzustellen. Die Berechnung des Korrelationskoeffizienten führen wir durch, um ein Maß über die Stärke des linearen Zusammenhangs der Merkmale zu erhalten.

2 Voraussetzung

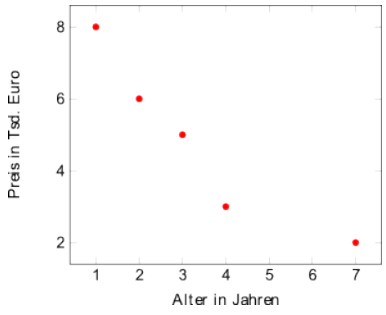
Voraussetzung für die Berechnung ist, dass es sich bei den Daten um mindestens zwei metrisch-skalierte Merkmale handelt, d.h. es handelt sich um quantitative Zahlenwerte, bei denen wir neben der Rangordnung und Häufigkeit auch das arithmetische Mittel bestimmen können.

3 Der Korrelationskoeffizient nach Pearson

Wir betrachten Alfreds Autohandel für den ein Kunde gerne den Zusammenhang zwischen dem Alter des Fahrzeugs und seinem Preis bei Wagen des gleichen Typs ermitteln möchte. Alfred hat folgende 5 Wagen zur Verfügung:

Wagen Nr. i	1	2	3	4	5
Alter in Jahren (x_i)	2	4	7	1	3
Preis in Tsd. Euro (y_i)	6	3	2	8	5

Zunächst einmal können wir uns die Tabelle als Diagramm veranschaulichen, sodass wir eine Tendenz erkennen können.



Da es sich beim *Alter des Autos* und dem *Preis* um metrisch-skalierte Merkmale handelt, können wir den Korrelationskoeffizienten berechnen, um Stärke und Richtung des Zusammenhanges anzugeben. Den Koeffizienten berechnen wir als das Verhältnis der Kovarianz beider Merkmale und dem Produkt der Standardabweichungen:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \qquad \text{Wertebereich: } -1 \leq \rho_{xy} \leq 1$$

Das Ergebnis für ρ_{xy} gibt uns die Stärke des linearen Zusammenhangs der Größen an. ρ_{xy} in der Nähe von -1 wird als stark negative Korrelation (dementsprechend als stark negativer, linearer Zusammenhang) und ρ_{xy} in der Nähe von 1 als stark positive Korrelation (dementsprechend als stark positiver, linearer Zusammenhang) bezeichnet. Hingegen besitzen die Merkmale für $\rho_{xy} = 0$ keinen linearen Zusammenhang.

3.1 Lösungsschritte

Den Korrelationskoeffizient können wir nach folgendem Muster berechnen:

1. Berechne die Kovarianz σ_{xy} oder auch als cov_{xy}
2. Berechne die Standardabweichungen σ_x und σ_y
3. Berechne den Korrelationskoeffizienten ρ_{xy}

Damit können wir nun die Ausgangsaufgabe lösen.

3.1.1 Kovarianz-Formel

Achtung bei der Formel der Kovarianz!

Wenn es **unterschiedliche Eintrittswahrscheinlichkeiten** für die Zustände gibt, wird folgende Formel benutzt:
$\sigma_{xy} = \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y) q_s$
q_s = Eintrittswahrscheinlichkeit des Zustandes s_i

Sind es **gleiche Eintrittswahrscheinlichkeiten**, dann wird einfach durch die Anzahl (n) der Zustände dividiert:
$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y)$

3.1.2 Formel der Standardabweichung

Achtung ebenfalls bei der Formel der Standardabweichung!

Wenn es **unterschiedliche Eintrittswahrscheinlichkeiten** für die Zustände gibt, wird folgende Formel benutzt:

$$\sigma_x = \sqrt{\sum_{i=1}^{n}(x_i - E_x)^2 q_s}$$

$$\sigma_y = \sqrt{\sum_{i=1}^{n}(y_i - E_y)^2 q_s}$$

Sind es **gleiche Eintrittswahrscheinlichkeiten**, dann wird einfach durch die Anzahl (n) der Zustände dividiert:

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - E_x)^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - E_y)^2}$$

3.1.3 Formel des Korrelationskoeffizienten

In die Formel des Korrelationskoeffizienten sind die Kovarianz $\sigma_x y$ und die Varianzen σ_x sowie σ_y einzusetzen.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \qquad \text{Wertebereich: } -1 \leq \rho_{xy} \leq 1$$

3.2 Lösung

1. Berechne die Kovarianz

Da keine Eintrittswahrscheinlichkeiten in der Aufgabe gegeben sind oder sich aus der Aufgabenstellung ableiten lassen, verwenden wir die Formel für **GLEICHE** Eintrittswahrscheinlichkeiten:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - E_x) \cdot (y_i - E_y)$$

2. Berechne die Standardabweichungen

Da keine Eintrittswahrscheinlichkeiten in der Aufgabe gegeben sind oder sich aus der Aufgabenstellung ableiten lassen, verwenden wir die Formel für **GLEICHE** Eintrittswahrscheinlichkeiten:

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y)^2}$$

Um die Kovarianz und die Standardabweichungen berechnen zu können benötigen wir einige Zwischenergebnisse, sodass es sich lohnt die Berechnung in Form einer Tabelle durchzuführen.

i	x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$	$(x_i - \mu_x) \cdot (y_i - \mu_y)$
1	2	6	−1,40	1,20	1,92	1,44	−1.68
2	4	3	0,60	−1,80	0,36	3,24	−1.08
3	7	2	3,60	−2,80	12,96	7,84	−10,08
4	1	8	−2,40	3,20	5,76	10,24	−7,68
5	3	5	−0,40	0,20	0,16	0.040	−0,08
Summe Σ	17	24	-	-	21,20	22,80	−20,60
$\frac{1}{n} \cdot \Sigma$	3,40	4,80	-	-	**4,24**	**4,56**	**−4,12**

Aus der Tabelle können wir nun folgende Ergebnisse entnehmen:
Kovarianz $\sigma_{xy} = -4{,}12$
Standardabweichung $\sigma_x = \sqrt{4{,}24}$
Standardabweichung $\sigma_y = \sqrt{4{,}56}$

3. Berechne den Korrelationskoeffizienten ρ_{xy}

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{-4{,}12}{\sqrt{4{,}24} \cdot \sqrt{4{,}56}} \approx -0{,}936983$$

Interpretation: Das Alter des Fahrzeugs und dessen Preis stehen in einem stark negativen, linearen Zusammenhang, d.h. je älter das Fahrzeug, desto niedriger der Preis.

4 Übungsaufgaben

1) Für 10 Filialen einer Handelskette soll untersucht werden, welcher Zusammenhang zwischen Verkaufsfläche (in m^2) und Umsatz (in Mio. Euro) besteht.

Filiale i	Fläche	Umsatz
1	150	3
2	180	8
3	420	19
4	480	22
5	660	31
6	1000	42
7	1300	48
8	1500	52
9	1600	54
10	1710	61

Berechne den Korrelationskoeffizienten nach Pearson.

Lösung

$\rho_{xy} \approx 0{,}986$

2) Zwei Wertpapiere WP_1 und WP_2 geben in den verschiedenen Zuständen s jeweils verschiedene Renditen r aus! Dabei treten die Zustände mit der Wahrscheinlichkeit q ein.

Ermittle den Korrelationskoeffizienten und interpretiere diesen!

	$s = 1$	$s = 2$
	$q = 0{,}40$	$q = 0{,}60$
WP_1	0,10	0,20
WP_2	0,30	0,00

Lösung

$Cov_{12} = -0{,}0072$
$\sigma_1 = 0{,}0024$
$\sigma_2 = 0{,}0216$
$\rho_{12} = -1$

Interpretation: Die Wertpapiere korrelieren sehr stark, so zusagen perfekt, negativ miteinander. Betrachtet man die Werte für r in den Zuständen s, dann steigen sie von $s = 1$ zu $s = 2$ für WP_1. Für WP_2 fallen sie von $s = 1$ zu $s = 2$. Die Renditen entwickeln sich gegenläufig. Tritt $s = 2$ ein, dann ist WP_1 besser, bei $s = 1$ ist WP_2 besser.