

Musterlösung Aufgabe 11: Korrelation und lineare Regression (per Hand)

Gegeben seien drei Beobachtungen eines Datensatzes mit zwei Variablen X und Y:

$$x_1 = 0, x_2 = 2, x_3 = -2, \quad y_1 = 5, y_2 = 4, y_3 = 3.$$

Führen Sie die folgenden Rechnungen per Hand durch.

- (a) Berechnen Sie für die beiden Variablen Mittelwert, Varianz und Standardabweichung.
- (b) Berechnen Sie für die beiden Variablen den Korrelationskoeffizienten nach Bravais-Pearson.
- (c) Berechnen Sie die Regressionsparameter des linearen Modells $y = c + dx$, bei dem also Y durch X vorhergesagt wird.

Lösung:

(a)

$$\begin{aligned}\bar{x} &= \frac{0 + 2 + (-2)}{3} = 0 \\ \bar{y} &= \frac{5 + 4 + 3}{3} = 4 \\ s_x^2 &= \frac{(0 - 0)^2 + (2 - 0)^2 + (-2 - 0)^2}{3 - 1} = \frac{0 + 4 + 4}{2} = 4 \Rightarrow s_x = 2 \\ s_y^2 &= \frac{(5 - 4)^2 + (4 - 4)^2 + (3 - 4)^2}{3 - 1} = \frac{1 + 0 + 1}{2} = 1 \Rightarrow s_y = 1\end{aligned}$$

(b)

$$\begin{aligned}s_{xy} &= \frac{(0 - 0)(5 - 4) + (2 - 0)(4 - 4) + (-2 - 0)(3 - 4)}{3 - 1} = -\frac{(-2)(-1)}{2} = 1 \\ r_{xy} &= \frac{s_{xy}}{s_x s_y} = \frac{1}{2 \cdot 1} = \frac{1}{2} = 0.5\end{aligned}$$

(c)

$$\begin{aligned}d &= \frac{s_{xy}}{s_x^2} = \frac{1}{4} \\ c &= \bar{y} - d\bar{x} = 4 - \frac{1}{4} \cdot 0 = 4 \\ \Rightarrow y &= c + dx = 4 + \frac{1}{4}x\end{aligned}$$

Zusatz: Es soll X durch Y vorhergesagt werden. Dann erhält man

$$\begin{aligned}d &= \frac{s_{xy}}{s_y^2} = \frac{1}{1} = 1 \\c &= \bar{x} - d\bar{y} = 0 - 1 \cdot 4 = -4 \\ \Rightarrow x &= c + dy = -4 + y\end{aligned}$$

Die letzte Gleichung entspricht $y = 4 + x$, d.h. man sieht, dass die beiden Geradengleichungen nicht genau übereinstimmen.

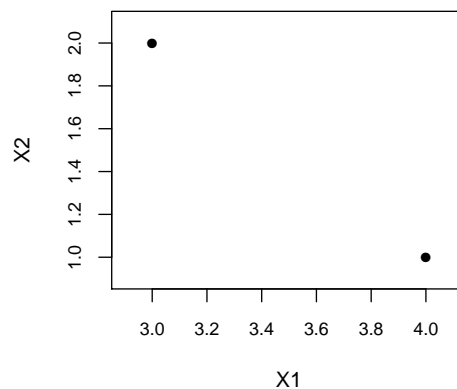
Musterlösung Aufgabe 12: Interpretation der Korrelation

Gegeben seien n Beobachtungen, für die jeweils die Werte für zwei Variablen X_1 und X_2 vorliegen. Welche der folgenden Aussagen sind richtig und welche sind falsch?

- a) Wenn X_1 immer größer als X_2 ist, dann ist der Korrelationskoeffizient (Pearson) positiv.
- b) Wenn die Werte von X_1 und X_2 in einem Streudiagramm abgetragen exakt auf einer Geraden liegen, genau dann ist der Betrag der Korrelationskoeffizienten (Pearson) exakt 1.
- c) Die Rangkorrelationskoeffizient (Spearman) von X_1 und $2X_2 + 3$ ist gleich dem Rangkorrelationskoeffizienten (Spearman) von X_1 und X_2 .
- d) Der Rangkorrelationskoeffizient (Spearman) ist mindestens so groß wie die Korrelationskoeffizient (Pearson).

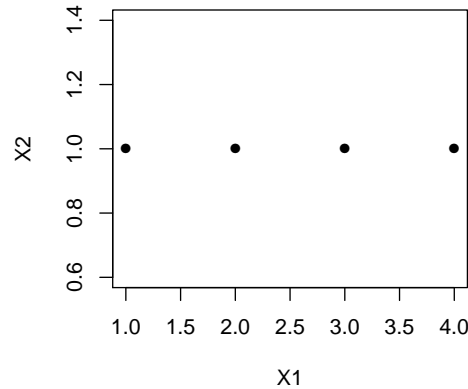
Lösung:

- (a) Die Aussage ist falsch. Hier ist ein Beispiel für $n = 2$ Beobachtungen. Für $x_{1,1} = 3$, $x_{1,2} = 4$ und $x_{2,1} = 2$, $x_{2,2} = 1$ ist der Korrelationskoeffizient -1 (Punkte liegen auf fallender Gerade), obwohl X_1 immer größer als X_2 ist.



- (b) Die Aussage ist falsch. Gilt beispielsweise $\text{Var}(X_2) = 0$, so ist der Korrelationskoeffizient nicht definiert:

$$\frac{s_{xy}}{s_x \cdot s_y} = \frac{0}{s_x \cdot 0}$$



- (c) Die Aussage ist richtig: Der Rangkorrelationskoeffizient kann berechnet werden, indem zunächst Ränge gebildet werden. Die Ränge der Beobachtungen X_2 und der transformierten Beobachtungen $2X_2 + 3$ sind dieselben, da die Funktion $x \mapsto 2x + 3$ eine streng monoton steigende Funktion ist.

Zusätzlich wird nun noch gezeigt, dass dieselbe Aussage auch für den Korrelationskoeffizienten (Pearson) gilt. Zunächst gilt für die (empirische) Kovarianz von $X := X_1$ und $Y := 2X_2 + 3$

$$\begin{aligned} s_{x_1, 2x_2+3} &= \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x}) (y_n - \bar{y}) \\ &= \frac{1}{N-1} \sum_{n=1}^N (x_{1,n} - \bar{x}_1) (2x_{2,n} + 3 - (2\bar{x}_2 + 3)) \\ &= 2 \frac{1}{N-1} \sum_{n=1}^N (x_{1,n} - \bar{x}_1) (x_{2,n} - \bar{x}_2) \\ &= 2 s_{x_1, x_2} \end{aligned}$$

Analog gilt für die Varianz von $Y = 2X_2 + 3$ und damit für die Standardabweichung:

$$\begin{aligned} s_y^2 &= s_{2x_2+3}^2 = 4 s_{x_2}^2 \\ s_y &= s_{2x_2+3} = 2 s_{x_2} \end{aligned}$$

Für die Korrelation erhält man somit

$$r_{x_1, 2x_2+3} = \frac{s_{x_1, 2x_2+3}}{s_{x_1} s_{2x_2+3}} = \frac{2 s_{x_1, x_2}}{s_{x_1} \cdot 2 s_{x_2}} = \frac{s_{x_1, x_2}}{s_{x_1} \cdot s_{x_2}} = r_{x_1, x_2}.$$

- (d) Die Aussage ist falsch, siehe z.B. Musterlösung zu Aufgabe 10 mit Rangkorrelationskoeffizient (Spearman) 0.952 und Korrelationskoeffizient (Pearson) 0.990.

Musterlösung Aufgabe 13: Wahrscheinlichkeitstheorie: Mengentheoretische Grundlagen

Ein Grundraum sei gegeben durch $\Omega = \{3, 3.1, \pi, 13, 33\}$.

- a) Welche Ergebnisse gehören zu den folgenden auf Ω eingeschränkten Ereignissen?
A: natürliche Zahlen; B: rationale Zahlen; C: Primzahlen.
- b) Wie sehen jeweils die paarweisen Schnittmengen und Vereinigungen der Ereignisse A, B und C^c (Komplement von C) aus?
- c) Wie sieht $B \setminus C$ aus und wie das Komplement von A?
- d) Angenommen, für die Wahrscheinlichkeiten, mit denen die Elementarereignisse auftreten, würde gelten:

$$P(\{3.1\}) = 0.35, \quad P(\{\pi\}) = 0.05, \quad P(\{3\}) = P(\{13\}) = P(\{33\}).$$

Berechnen Sie für alle Mengen aus den Aufgabenteilen b) und c) deren Wahrscheinlichkeiten.

Lösung:

- (a) $A = \{3, 13, 33\}$;
 $B = \{3, 3.1, 13, 33\}$;
 $C = \{3, 13\}$.
- (b) $C^c = \Omega \setminus C = \{3.1, \pi, 33\}$

	Schnitt	Vereinigung
A mit B	$\{3, 13, 33\}$	$\{3, 3.1, 13, 33\}$
A mit C^c	$\{33\}$	$\{3, 3.1, \pi, 13, 33\}$
B mit C^c	$\{3.1, 33\}$	$\{3, 3.1, \pi, 13, 33\}$

- (c) $B \setminus C = \{3.1, 33\}$
 $A^c = \Omega \setminus A = \{3.1, \pi\}$

- (d) Zunächst gilt: $P(\{3\}) = P(\{13\}) = P(\{33\}) = (1 - 0.35 - 0.05)/3 = 0.2$.

	Schnitt	Vereinigung
A mit B	$3 \cdot 0.2 = 0.6$	$1 - 0.05 = 0.95$
A mit C^c	0.2	1
B mit C^c	$0.35 + 0.2 = 0.55$	1

$$P(B \setminus C) = P(\{3.1, 33\}) = 0.35 + 0.2 = 0.55 \quad \text{oder} \quad P(B \setminus C) = P(B \cap C^c) = 0.55$$

$$P(A^c) = P(\{3.1, \pi\}) = 0.35 + 0.05 = 0.4$$

Musterlösung Aufgabe 14: Wahrscheinlichkeitstheorie: Regeln für Wahrscheinlichkeiten

Gegeben sei ein Wahrscheinlichkeitsraum (Ω, \mathcal{A}, P) .

- a) Wann gilt $P(A \cup B) = P(A) + P(B)$ und wann gilt $P(A \cup B) > P(A) + P(B)$?
- b) Welche Wahrscheinlichkeit ist größer, $P(A \cap B)$ oder $P(A) \cdot P(B)$?
- c) Warum gilt für Wahrscheinlichkeiten stets $P(A) \geq 0$ und $P(A) \leq 1$?

Lösung:

- (a) Laut Eigenschaft (iii) von Wahrscheinlichkeitsmaßen (Kapitel 4 der Vorlesung, Folie 61) gilt:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Daher gilt $P(A \cup B) = P(A) + P(B)$ genau dann, wenn $P(A \cap B) = 0$ ist. Wegen $P(A \cap B) \geq 0$ kann $P(A \cup B) > P(A) + P(B)$ nie gelten.

- (b) Beide Wahrscheinlichkeiten können größer sein. Für Gleichheit gilt gerade Unabhängigkeit der Ereignisse (später in der Vorlesung).

Für $A = B$ mit $0 < P(A) < 1$ gilt:

$$P(A \cap B) = P(A) > P(A) \cdot P(A) = P(A) \cdot P(B).$$

Falls $A \cap B = \emptyset$ mit $P(A) > 0$ und $P(B) > 0$ gilt:

$$P(A \cap B) = P(\emptyset) = 0 < P(A) \cdot P(B).$$

- (c) Dies ist das erste Kolmogorov-Axiom für Wahrscheinlichkeitsmaße (Kapitel 4 der Vorlesung, Folie 46).