FH GR Fachhochschule Graubünden
University of Applied Sciences

Übungsblatt Sto 9

Computational and Data Science
BSc HS2024

Lösungen

Mathematik 3

Lernziele:
- Sie kennen die Begriffe Punktwolke, Kovarianz, Korrelationskoeffizient, Bestimmtheitsmass und ihre wichtigsten Eigenschaften.
- Sie können die Kovarianz und Korrelationskoeffizienten für eine gegebene Stichprobe bestimmen und können die Korrelation von Daten mittels einer Punktwolke qualitativ beurteilen.

---

1. Kovarianz und Korrelationskoeffizient

Berechnen Sie die Kovarianz und den Korrelationskoeffizienten der folgenden Stichproben und skizzieren Sie die jeweilige Punktwolke:

a)

i	1	2	3	4	5	6
x_i	2	2	4	1	5	4
y_i	3	2	4	2	4	2

b)

i	1	2	3	4	5
x_i	2	4	6	8	10
y_i	20	17	13	10	6

a)

i	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	3	-1	1	1/6	1/36	$-1/6$
2	2	2	-1	1	$-5/6$	25/36	5/6
3	4	4	1	1	7/6	49/36	7/6
4	1	2	-2	4	$-5/6$	25/36	10/6
5	5	4	2	4	7/6	49/36	14/6
6	4	2	1	1	$-5/6$	25/36	$-5/6$
$\sum$	18	17	0	12	0	174/36	5

$$\bar{x} = \frac{1}{6} \cdot \sum_{1}^{5} x_i = \frac{1}{6} \cdot 18 = 3; \qquad \bar{y} = \frac{1}{6} \cdot \sum_{1}^{5} y_i = \frac{1}{6} \cdot 17 = \frac{17}{6}$$

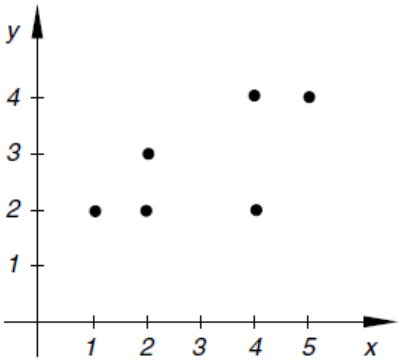
Summation geht jeweils bis n = 6.

$$s_x^2 = \frac{1}{6-1} \cdot \sum_{1}^{5} (x_i - \bar{x})^2 = \frac{1}{5} \cdot 12 = 2{,}4 \quad \Rightarrow \quad s_x = 1{,}5492$$

$$s_y^2 = \frac{1}{6-1} \cdot \sum_{1}^{5} (y_i - \bar{y})^2 = \frac{1}{5} \cdot \frac{174}{36} = 0{,}9667 \quad \Rightarrow \quad s_y = 0{,}9832$$

$$s_{xy} = \frac{1}{6-1} \cdot \sum_{1}^{5} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{5} \cdot 5 = 1$$

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{1}{1{,}5492 \cdot 0{,}9832} = 0{,}6565$$



b) Tabelle wie in a) erstellen

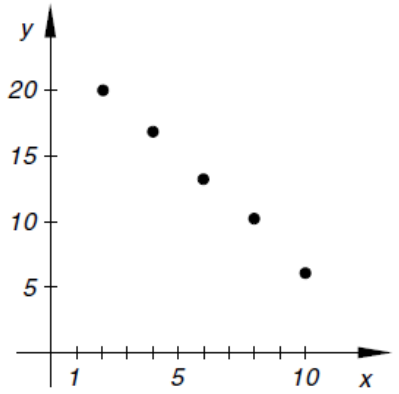
$$\bar{x} = \frac{1}{5} \cdot \sum_{1}^{5} x_i = \frac{1}{5} \cdot 30 = 6; \qquad \bar{y} = \frac{1}{5} \cdot \sum_{1}^{5} y_i = \frac{1}{5} \cdot 66 = 13{,}2$$

$$s_x^2 = \frac{1}{5-1} \cdot \sum_{1}^{5} (x_i - \bar{x})^2 = \frac{1}{4} \cdot 40 = 10; \qquad s_x = \sqrt{10}$$

$$s_y^2 = \frac{1}{5-1} \cdot \sum_{1}^{5} (y_i - \bar{y})^2 = \frac{1}{4} \cdot 122{,}8 = 30{,}7; \qquad s_y = 5{,}5408$$

$$s_{xy} = \frac{1}{5-1} \cdot \sum_{1}^{5} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{4} \cdot (-70) = -17{,}5$$

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{-17{,}5}{\sqrt{10} \cdot 5{,}5408} = -0{,}9988$$



2. Körpergrösse und -gewicht von Studierenden

Berechnen Sie den empirischen Korrelationskoeffizient zwischen Körpergrösse und Körpergewicht für die Werte von 10 Studierenden und fertigen Sie eine Punktwolke für die Werte an.

Körpergrösse (m)	1,8	1,75	1,83	1,65	1,77	1,73	1,79	1,69	1,90	1,81
Körpergewicht (kg)	84	70	95	72	65	82	72	68	95	80

Die Mittelwerte bzw. empirischen Standardabweichungen der Körpergrößen und Körpergewichte sind gegeben durch

$$\bar{x} = \frac{1{,}80 + 1{,}75 + 1{,}83 + \cdots + 1{,}81}{10} = 1{,}77$$

$$\bar{y} = \frac{84 + 70 + 95 + \cdots + 80}{10} = 78{,}3$$

$$s_x = \sqrt{\frac{1}{9}((1{,}80 - 1{,}77)^2 + (1{,}75 - 1{,}77)^2 + \cdots + (1{,}81 - 1{,}77)^2)} = 0{,}0717$$

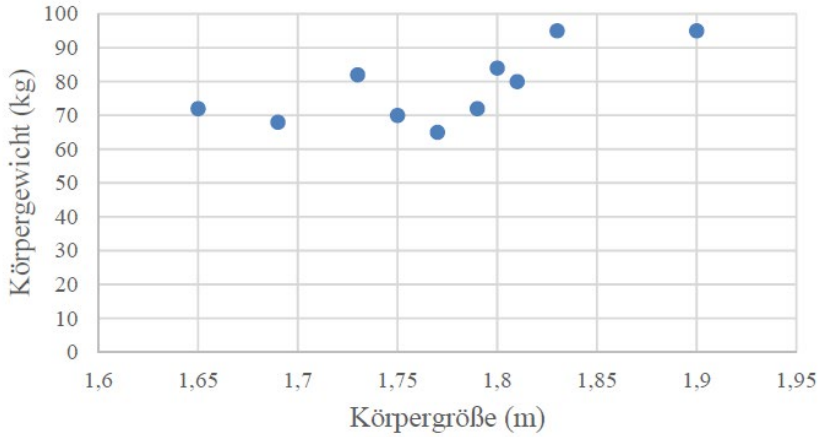
$$s_y = \sqrt{\frac{1}{9}((84 - 78{,}3)^2 + (70 - 78{,}3)^2 + \cdots + (80 - 78{,}3)^2)} = 10{,}7399$$

Die empirische Kovarianz zwischen den Körpergrößen und Körpergewichten errechnet sich aus

$$s_{xy} = \frac{1}{10-1}\left(\sum_{i=1}^{10} x_i y_i - 10\bar{x}\bar{y}\right)$$
$$= \frac{1}{9}\big((1{,}80)(84) + (1{,}75)(70) + \cdots + (1{,}81)(80)$$
$$- 10(1{,}77)(78{,}3)\big) = 0{,}5427$$

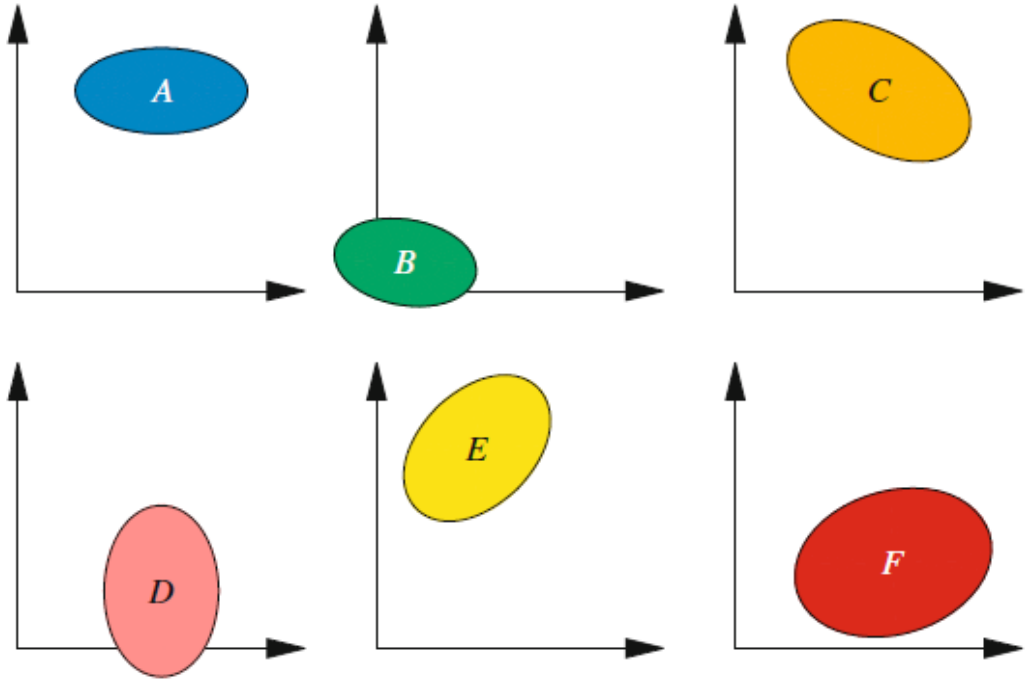
Folglich ermittelt man den empirischen Korrelationskoeffizient durch

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{0{,}5427}{(0{,}0717)(10{,}7399)} = 0{,}7048$$



3. Korrelationen bei Punktwolken

In der Abbildung sind sechs verschiedene Punktwolken symbolisch durch Ellipsen angezeigt. In welchen Punktwolken ist die Korrelation positiv oder negativ? Wo ist die Korrelation gleich null? Ordnen Sie die Punktwolken nach der Größe ihrer Korrelation von -1 bis +1.



Bezeichen wir mit r_A die Korrelation der Punktwolke A und entsprechend auch die der anderen Punktwolken, so gilt:

$$r_C < r_B < 0 \approx r_A \approx r_D < r_F < r_E$$

4. Korrelationen

a) Bei einer Verpackungsmaschine seien das Nettogewicht *N* des Füllgutes und das Gewicht *T* der Verpackung voneinander unabhängig. Das Bruttogewicht *B* ist die Summe aus beiden
B = *N* + *T*. Sind *B* und *N* unkorreliert oder positiv- oder negativ-korreliert?

b) Sei E_S die von der Sonne eingestrahlte und E_P die von Pflanzen genutzte Energie. Hängt dann die Korrelation zwischen E_S und E_P davon ab, ob E_S durch die Wellenlänge oder die Frequenz c des Lichtes gemessen wird?

c) Unterstellt man feste Umrechnungskurse zwischen den nationalen Währungen und dem Euro, sind dann die Korrelationen zwischen Einfuhr- und Ausfuhrpreisen abhängig davon, ob die Preise in DM oder in Euro gemessen werden?

a) B und N sind positiv korreliert, denn cov(B,N) = cov(N + T,N) = cov(N,N)+cov(T,N) = var(N) > 0, da cov(N,N) = var(N) und cov(T,N) = 0, da T und N unkorreliert sind.

b) Ja, denn die Beziehung zwischen Wellenlänge λ und Frequenz f ist nicht linear ($f = c/\lambda$ mit c = Lichtgeschwindigkeit).

c) Nein, denn der Umrechnungskurs von Euro in DM ist linear.

5. Störche und Wasserfläche

Angenommen, wir haben die in der folgenden Tabelle dargestellten Daten aus 10 Ländern vorliegen: Dabei bedeuten F_i die Fläche des i-ten Landes in km^2, B_i die

Anzahl (in Tausend) der im letzten Jahr dort geborenen Babys, S_i die Anzahl der Störche und W_i die Gesamtgrösse der Wasserfläche in km^2 (die Zahlen sind fiktiv).

					Quoten mal 100			
Land	F_i	B_i	S_i	W_i	S_F	B_F	S_W	
1	8624	370	213	157	2,47	4,3	135	
2	9936	210	48	150	0,48	2,1	32	
3	2093	323	100	190	4,78	15,4	53	
4	3150	306	152	185	4,83	9,7	82	
5	4584	373	146	177	3,18	8,1	82	
6	4294	556	95	179	2,21	13,0	53	
7	15570	520	85	122	0,55	3,3	69	
8	9260	300	149	154	1,61	3,2	97	
9	2377	580	149	288	6,27	24,4	52	
10	12149	287	192	139	1,58	2,4	138	

a) Bestimmen Sie die Korrelationen r(F,B), r(F,S) und r(B,S).
b) Beziehen Sie dann die jeweilige Anzahl der Babys und der Störche auf die zur Verfügung stehende Fläche. Es sei $(S_F)_i = S_i/F_i$ und $(B_F)_i = B_i/F_i$ die Anzahl der Störche pro km^2 bzw. die Anzahl der Babys pro km^2. Bestimmen Sie die Korrelation $r(S_F,B_F)$.
c) Nun beziehen Sie die Anzahl der Störche nicht auf die Grösse des Landes, sondern auf die Größe der nahrungsspendenden Wasserfläche W. Jetzt ist $(S_W)_i = S_i/W_i$ die Anzahl der Störche pro km^2 Wasserfläche. Bestimmen Sie die Korrelation $r(S_W,B_F)$.

Was lässt sich aus diesen Korrelationen lernen?

Wir berechnen die Korrelation $r(F, B)$ zwischen den Merkmalen F und B. Die folgende Tabelle zeigt alle notwendigen Rechenschritte. Dabei sind alle Zahlen auf ganze Zahlen gerundet angegeben. Gerechnet wurde aber mit 4 Dezimalstellen nach dem Komma.

i	F_i	B_i	$(F_i - \overline{F})^2$	$(B_i - \overline{B})^2$	$(F_i - \overline{F})(B_i - \overline{B})$
1	8624	370	2 017 252	156	−17 754
2	9936	210	7 465 463	29 756	−471 322
3	2093	323	26 119 254	3540	304 087
4	3150	306	16 432 484	5852	310 108
5	4584	373	6 862 828	90	24 887
6	4294	556	8 466 354	30 102	−504 833
7	15 570	520	69 994 976	18 906	1 150 366
8	9260	300	4 228 370	6806	−169 645
9	2377	580	23 297 033	39 006	−953 273
10	12 149	287	24 455 992	9120	−472 276
$\sum$	72 037	3825	189 340 006	143 337	−799 655

Daraus folgt $\overline{F} = \frac{72\,037}{10} = 7203.7$, $\overline{B} = 382.5$, $\text{var}(\boldsymbol{F}) = 18\,934\,000.6$, $\text{var}(\boldsymbol{B}) = 14\,333.7$ und $\text{cov}(\boldsymbol{F}, \boldsymbol{B}) = -79\,965.5$. Dann ist

$$r(F, B) = \frac{\text{cov}(\boldsymbol{F}, \boldsymbol{B})}{\sqrt{\text{var}(\boldsymbol{F}) \cdot \text{var}(\boldsymbol{B})}} = \frac{-79\,965.5}{\sqrt{18\,934\,000.6 \cdot 14\,333.7}} = -0.154$$

Für die weiteren Schritte geben wir die Kovarianzmatrix an, wie oben ganzzahlig gerundet:

cov	F	B	S	S_F	B_F	S_W
F	18 934 001	−79 965	−10 461	−6849	−23 694	51 507
B	−79 965	14 334	−294	62	488	−964
S	−10 461	−294	2298	24	−11	1423
S_F	−6849	62	24	3	11	−10
B_F	−23 694	488	−11	11	47	−106
S_W	51 507	−964	1423	−10	−106	1135

Auf den Diagonalen dieser Matrix stehen die Merkmale. Wie bei der Berechnung von r(F,B) bestimmt man nun die paarweisen Korrelationen.

r	F	B	S	S_F	B_F	S_W
F	1	−0.15	−0.05	−0.85	−0.79	0.35
B	**−0.15**	1	−0.05	0.28	0.59	−0.24
S	**−0.05**	**−0.05**	1	0.28	−0.03	0.88
S_F	−0.85	0.28	0.28	1	0.87	−0.17
B_F	−0.79	0.59	−0.03	**0.87**	1	−0.46
S_W	0.35	−0.24	0.88	−0.17	**−0.46**	1

B, S und F sind praktisch unkorreliert: $r(F, B) = -0.15$, $r(F, S) = -0.05$ und $r(B, S) = -0.05$.

Trotzdem sind die Baby- und die Storchquoten hochgradig miteinander korreliert $r(B_F, S_F) = 0.87$! Ein arg naiver Betrachter könnte aus der erfreulichen Zunahme der Störche auch auf ein Ansteigen der Geburtenziffer hoffen. Berücksichtigt man dagegen die Wasserfläche, so sind auf einmal Baby- und Storchquoten negativ korreliert: $r(B_F, S_W) = -0.46$! Also je weniger Störche, umso mehr Kinder! Oder??

Was lässt sich aus diesem Beispiel lernen? Eine kausale Verknüpfung zwischen F, B und S ist nirgends ersichtlich. Berechnet man jedoch Quoten oder andere Gliederungszahlen, so können sich die Korrelationen unvorhersehbar ändern.