Stem-and-leaf-plot

– Spezielles Histogramm mit Klassen nach Dezimalsystem und reproduzierbaren Einzeldaten

– Vorteil: trotz Gruppierung kein Informationsverlust

– Beispiel: Alter von Chefredakteuren

3:	678	→ Die jüngsten drei Chefredakteure sind 36, 37 und 38
4:	00111112222222233444	
4:	56667788888999	
5:	000000112222224444	
5:	67888899	
6:	00000112234	

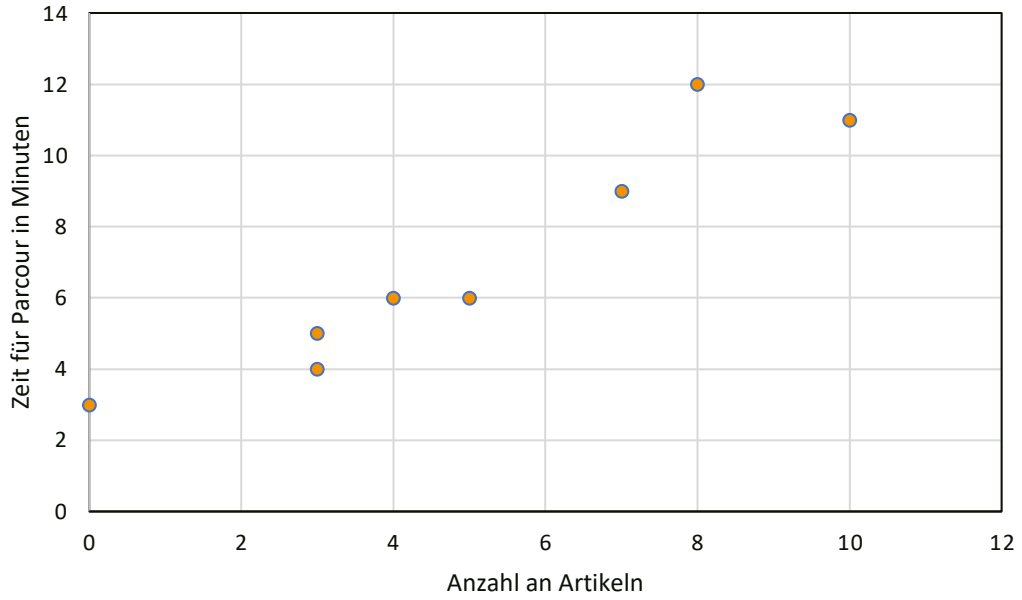
Beispiel für Datenanalyse

- 8 Personen verbringen einen Abend zusammen in der Bar, die Variable B gibt an, wie viele Biere sie am Ende insgesamt getrunken haben.
- Abschließend sollen alle 8 einen Geschicklichkeitsparkour durchlaufen. Die Variable P gibt an, wie lange sie für diesen in Minuten brauchen.

Person (n = 8)	Roman	Andi	Tom	Anika	Tanja	Svenja	Lukas	Maxi
B	8	0	5	3	3	10	4	7
P	12	3	6	5	4	11	6	9

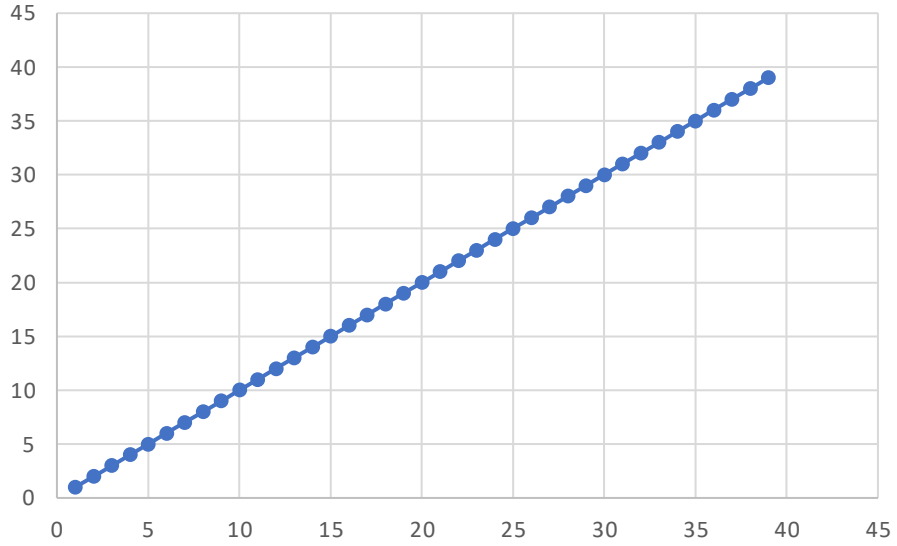
Steudiagramm/Scatterplot

- Abbildung eines Zusammenhangs zweier Merkmale durch eine Punktewolke
- Merkmale werden auf Achsen abgetragen
  - x-Achse/Abszisse: beeinflussendes Merkmal
  - y-Achse/Ordinate: abhängiges Merkmal (Zielgröße)

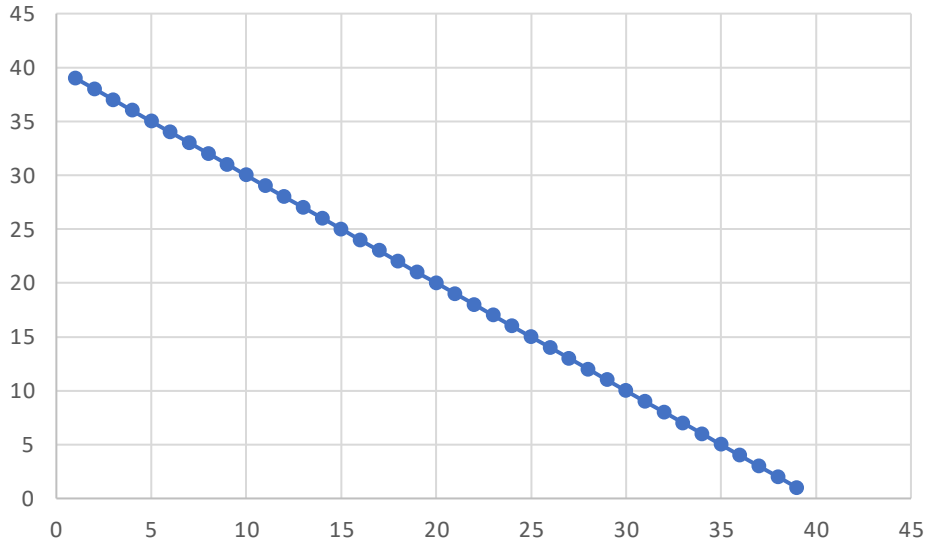


Steudiagramm/Scatterplot

Perfekte positive Korrelation (r = 1)



Perfekte negative Korrelation (r = -1)



Mehrdimensionale Analysen

– Beschreibung der Zusammenhänge in Abhängigkeit vom Skalenniveau

Metrische Untersuchungsmerkmale	Korrelationskoeffizient von Bravais-Pearson und Regressionsmodelle
Ordinale Merkmale	Rangkorrelationskoeffizient nach Spearman
Nominale Merkmale	Kontingenztabellen und Kontingenzkoeffizienten

Kovarianz

– Maß für den Zusammenhang zwischen zwei Merkmalen

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$\bar{B} = \frac{1}{8} \times (8 + 0 + 5 + 3 + 3 + 10 + 4 + 7) = 5$

$\bar{P} = \frac{1}{8} \times (12 + 3 + 6 + 5 + 4 + 11 + 6 + 9) = 7$

$S(BP) = \frac{1}{7} \times ((8-5) \times (12-7) + (0-5) \times (3-7) + (5-5) \times (6-7) + (3-5) \times (5-7) + (3-5) \times (4-7) + (10-5) \times (11-7) + (4-5) \times (6-7) + (7-5) \times (9-7))$

$= \frac{1}{7} \times (15 + 20 + 0 + 4 + 6 + 20 + 1 + 4)$

$= \frac{1}{7} \times 70$

$= 10$

Bravais-Pearson-KorKoef.

– Maß für den linearen Zusammenhang zweier Merkmale

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{mit } -1 \le r_{XY} \le +1$$

$S^2(B) = \frac{1}{7} \times (8 - 5)^2 + (0 - 5)^2 + (5 - 5)^2 + (3 - 5)^2 + (3 - 5)^2 + (10 - 5)^2 + (4 - 5)^2 + (7 - 5)^2$

$= \frac{1}{7} \times (9 + 25 + 0 + 4 + 4 + 25 + 1 + 4) \quad = \frac{1}{7} \times 72 \quad = 10{,}29$

S (B) = = **3,21**

$S^2(P) = \frac{1}{7} \times (12 - 7)^2 + (3 - 7)^2 + (6 - 7)^2 + (5 - 7)^2 + (4 - 7)^2 + (11 - 7)^2 + (6 - 7)^2 + (9 - 7)^2$

$= \frac{1}{7} \times (25 + 16 + 1 + 4 + 9 + 16 + 1 + 2) \quad = \frac{1}{7} \times 73 \quad = 10{,}43$

S (P) = = **3,23**

$$r_{BP} = \frac{S\,(BP)}{S\,(B)\,S\,(Y)} = \frac{10}{3{,}21 \times 3{,}23} = \mathbf{0{,}96}$$

Bravais-Pearson-KorKoef.

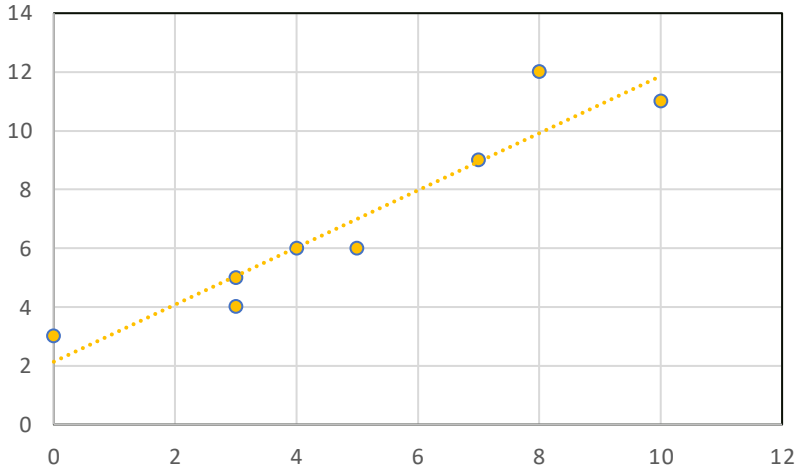
– Wie lässt sich der Wert für den Korrelationskoeffizienten interpretieren?

$r_{xy} > 0$	positiv korreliert
$r_{xy} = 0$	unkorreliert
$r_{xy} < 0$	negativ korreliert

$r_{xy} > \pm 0{,}8$	starker positiver/negativer Zusammenhang
$\pm 0{,}5 \leq r_{xy} \leq \pm 0{,}8$	mittlerer Zusammenhang
$r_{xy} < \pm 0{,}5$	schwacher positiver/negativer Zusammenhang
$r_{xy} = 0$	kein (linearer) Zusammenhang

Lineare Regression/KQ-Methode

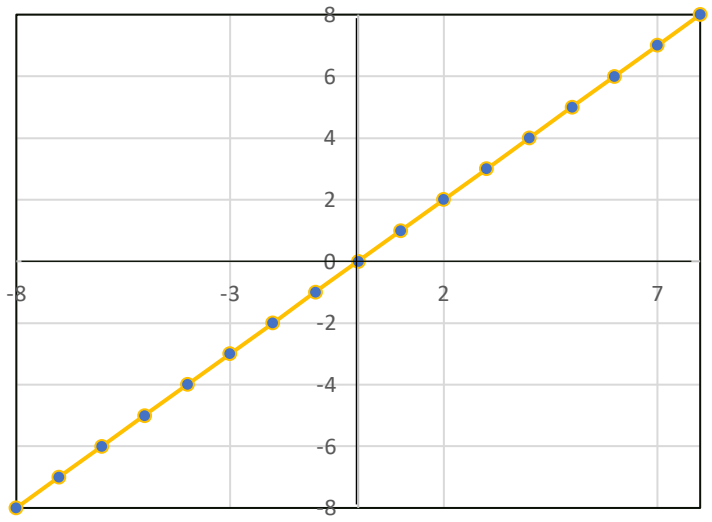
- Der Korrelationskoeffizient gibt uns Aufschluss darüber, ob eine Korrelation vorliegt, jedoch nicht in welche Richtung
- Ziel: Merkmal Y durch ein anderes Merkmal X erklären
- Nach dem Einzeichnen einer Gerade, kann für jeden Wert von x ein geschätzter y-Wert abgelesen werden
- Diesen nennt man $\hat{Y}$



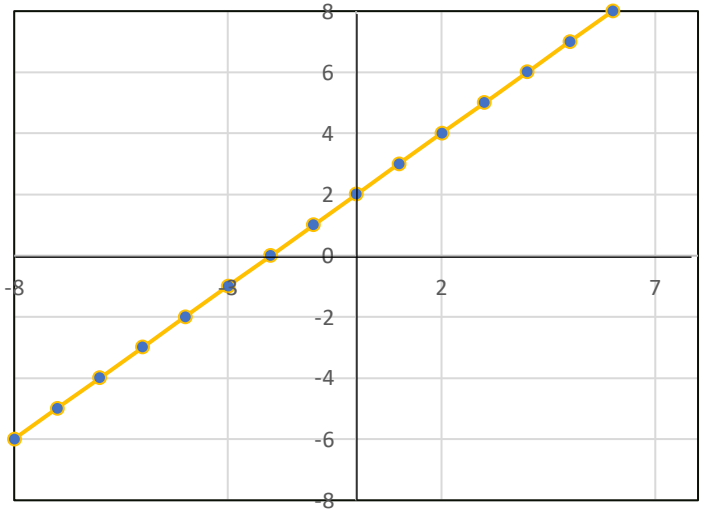
Exkurs: lineare Funktionen

$$F(x) = b\,x + a$$

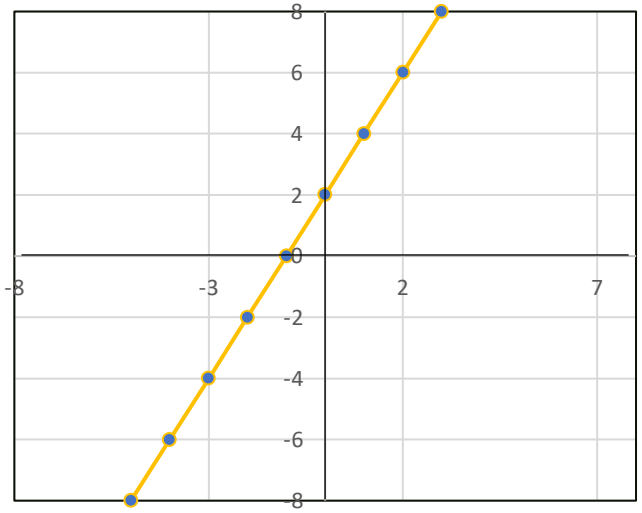
a = 0
b = 1



a = 2
b = 1



a = 2
b = 2



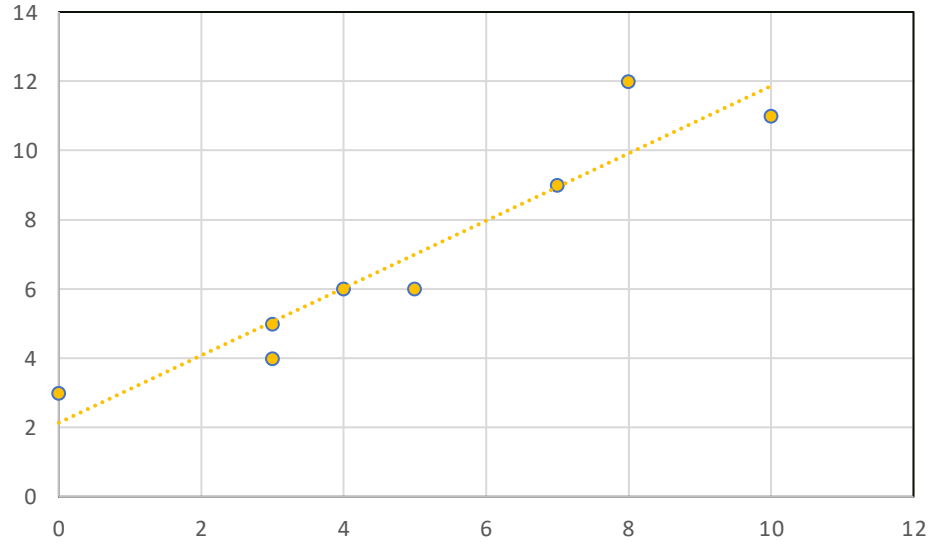
$$\rightarrow \hat{y}_i = a + b \cdot x_i$$

Lineare Regression/KQ-Methode

- Die Striche im Diagramm nennt man Residuen
- Summe der quadratischen Residuen soll minimal werden
- Sie werden minimal für:

$$b = \frac{s_{xy}}{s_x^2}$$

$$a = \bar{y} - b\bar{x}$$



Lineare Regression/KQ-Methode

$$b = \frac{s_{BP}}{s_B^2} = \frac{10}{10{,}29} = 0{,}97$$

$$a = \bar{P} - b \times \bar{B} = 7 - 0{,}97 \times 5 = 2{,}15$$

$$\hat{y}_i = a + b \cdot x_i = \mathbf{2{,}15 + 0{,}97\ x_i}$$

Regressionsgerade

– Eigenschaften der Regressionsgerade

Maßzahl	Kovarianz	Streudiagramm und Regressionsgeraden
$r = 1$	$s_{XY} > 0$	Alle Punkte des Streudiagramms liegen auf einer Geraden mit positiver Steigung. Beide Regressionsgeraden fallen zusammen.
$0 < r < 1$	$s_{XY} > 0$	Beide Regressionsgeraden haben eine positive Steigung. Das Streuungsdiagramm ist oft ellipsenförmig mit steigender Tendenz.
$r = 0$	$s_{XY} = 0$	Beide Regressionsgeraden sind achsenparallel und stehen im Punkt $(\bar{x}, \bar{y})$ senkrecht aufeinander.
$0 > r > -1$	$s_{XY} < 0$	Beide Regressionsgeraden haben eine negative Steigung. Das Streuungsdiagramm ist oft ellipsenförmig mit fallender Tendenz.
$r = -1$	$s_{XY} < 0$	Alle Punkte des Streudiagramms liegen auf einer Geraden mit negativer Steigung. Beide Regressionsgeraden fallen zusammen.

Uebungsaufgaben

Aufgabe 1

– Gegeben sind verschiedene Waschmaschinen mit Angabe der Anschaffungskosten (X) und des Stromverbrauchs (Y)

Waschmaschinen	Anschaffungskosten in € (x)	Stromverbrauch in kWh (y)
1	700	60
2	500	85
3	650	70
4	400	100
5	320	140

– Berechne ein geeignetes Maß für den Zusammenhang zwischen Anschaffungskosten und Stromverbrauch und interpretiere dieses.
– Bestimme die Gleichung der Regressionsgerade.

Aufgabe 1 - Lösung

$\overline{X} = \frac{1}{5} \times (700 + 500 + 650 + 400 + 320) = \frac{1}{5} \times 2570 = 514$

$\overline{y} = \frac{1}{5} \times (60 + 85 + 70 + 100 + 140) = \frac{1}{5} \times 455 = 91$

$S_{xy} = \frac{1}{4} \times ((700 - 514) \times (60 - 91) + (500 - 514) \times (85 - 91) + (650 - 514) \times (70 - 91) + (400 - 514) \times (100 - 91) + (320 - 514) \times (140 - 91) = -4767{,}5$

$S^2(x) = \frac{1}{4} \times (700 - 514)^2 + (500 - 514)^2 + (650 - 514)^2 + (400 - 514)^2 + (320 - 514)^2 = 25980$

$S^2(y) = \frac{1}{4} \times (60 - 91)^2 + (85 - 91)^2 + (70 - 91)^2 + (100 - 91)^2 + (140 - 91)^2 = 980$

$r_{xy} = \frac{4767}{\sqrt{25980} \times \sqrt{980}} = -0{,}94$ => starker negativer Zusammenhang

$b = \frac{-4767{,}5}{25980} = -0{,}18$

$a = 91 - (-0{,}18) \times 514 = 183{,}52$

$\hat{Y} = 183{,}52 - 0{,}18x$

Aufgabe 3.1 der AS

Aufgabe 3.1

Eine Befragung von 20 Haushalten über die Wohnungsgröße (Anzahl der Zimmer: X) un den Computer-Bestand (Anzahl der Computer: Y) ergab folgende Paare (x_i, y_i):

(1, 0)	(2, 2)	(3, 0)	(1, 2)	(2, 0)	(2, 0)	(3, 2)	(1, 1)	(2, 2)	(2, 1)
(1, 0)	(3, 2)	(2, 2)	(2, 0)	(3, 0)	(2, 2)	(1, 1)	(1, 0)	(3, 2)	(3, 1)

a) Ermitteln sie die zweidimensionale absolute und relative Häufigkeitsverteilung und geben sie dabei die Rangverteilung der beiden Merkmale.
b) Wie viele Zimmer und wie viele Computer sind im Durchschnitt je Haushalt vorhanden? Geben sie die entsprechenden (empirischen) Varianzen und (empirischen) Standardabweichungen an.
c) Wie groß ist der Anteil der Haushalte, bei dem höchstens zwei Zimmer und höchstens ein Computer vorhanden sind?
d) Wie viel Prozent der Haushalte mit einem Zimmer besitzen mind. einen Computer?
e) Wie viel Prozent der Haushalte, die mindestens einen Computer besitzen, haben eine Wohnung mit mindestens zwei Zimmern.
f) Berechnen und interpretieren Sie die Kovarianz und den Korrelationskoeffizienten zwischen Wohnungsgröße und Computer.

Aufgabe 3.1 der AS - Lösung (1/4)

a)

- Zweidimensionale absolute Häufigkeitsverteilung (X_i, Y_j) -> $h_{ij} = n_{ij}$

$x_i \setminus y_j$	0	1	2	Σ
1	3	2	1	6
2	3	1	4	8
3	2	1	3	6
Σ	8	4	8	20

- Zweidimensionale relative Häufigkeitsverteilung (X_i, Y_j) -> $r_{ij} = \frac{n_{ij}}{n}$

$x_i \setminus y_j$	0	1	2	$r_{i\cdot}$
1	0,15	0,10	0,05	0,3
2	0,15	0,05	0,20	0,4
3	0,10	0,05	0,15	0,3
$r_{\cdot j}$	0,4	0,20	0,4	1

b)

- $\bar{x} = \sum_{i=1}^{n} r_{i\cdot} \cdot x_i = 0{,}3 \cdot 1 + 0{,}4 \cdot 2 + 0{,}3 \cdot 3 = 2$ [Zimmer]
- $\bar{y} = \sum_{j=1}^{n} r_{\cdot j} \cdot y_j = 0{,}4 \cdot 0 + 0{,}2 \cdot 1 + 0{,}4 \cdot 2 = 1$ [Computer]
- $S_X^2 = \frac{1}{n\text{-}1} \sum_{i=1}^{n} n_i \cdot (x_i - \bar{x})^2 = \frac{1}{20\text{-}1} \{6 \cdot (1-2)^2 + 8 \cdot (2-2)^2 + 6 \cdot (3-2)^2\} = \frac{12}{19} \approx 0{,}63$ [Zimmer²]
- $S_x = \sqrt{S_X^2} \approx \sqrt{0{,}63} \approx 0{,}79$ [Zimmer]
- $S_Y^2 = \frac{1}{n\text{-}1} \sum_{j=1}^{n} n_j \cdot (y_j - \bar{y})^2 = \frac{1}{20\text{-}1} \{8 \cdot (0-1)^2 + 4 \cdot (1-1)^2 + 8 \cdot (2-1)^2\} = \frac{16}{19} \approx 0{,}84$ [Computer²]
- $S_Y = \sqrt{S_Y^2} \approx \sqrt{0{,}84} \approx 0{,}92$ [Computer]

Aufgabe 3.1 der AS - Lösung (2/4)

c)

Gesucht wird derjenige Anteil, der Haushalte r_H, bei denen gilt: $X_i \leq 2$ und gleichzeitig $Y_j \leq 1$

$x_i \setminus y_j$	0	1	2	$r_{i\cdot}$
1	0,15	0,10	0,05	0,3
2	0,15	0,05	0,20	0,4
3	0,10	0,05	0,15	0,3
$r_{\cdot j}$	0,4	0,20	0,4	1

- $R_H = 0{,}15 + 0{,}10 + 0{,}15 + 0{,}05 = 0{,}45$

d)

Betrachtet werden jetzt nur noch die Haushalte mit einem Zimmer.

$x_i = 1 \setminus y_j$	0	1	2	Σ
n_{1j}	3	2	1	6
r_{1j}	0,5	$2/6 \approx 0{,}33$	$1/6 \approx 0{,}17$	1

- $r_{H1} = 2/6 + 1/6 \quad = 0{,}5$ -> P(H1) = 50%

Aufgabe 3.1 der AS - Lösung (3/4)

e)
Jetzt werden nur noch die Haushalte mit mindestens einem Computer betrachtet. Die neue Basismenge sind demnach 12 Haushalte.

- Zweidimensionale absolute Häufigkeitsverteilung (X_i,Y_j) -> $h_{ij} = n_{ij}$

$x_i \setminus y_j$	1	2	Σ
1	2	1	3
2	1	4	5
3	1	3	4
Σ	4	8	12

- Zweidimensionale relative Häufigkeitsverteilung (X_i,Y_j) -> $r_{ij} = \frac{n_{ij}}{n}$

$x_i \setminus y_j$	1	2	$r_{i\cdot}$
1	2/12	1/12	3/12
2	1/12	4/12	5/12
3	1/12	3/12	4/12
$r_{\cdot j}$	4/12	8/12	1

- R = 1/12 + 4/12 + 1/12 + 3/12 = 9/12 = 0,75 -> $P(X \geq 2) = 75\%$

Aufgabe 3.1 der AS - Lösung (4/4)

f)

Die Kovarianz ist ein Streuungsmaß für zweidimensionale Verteilungen mit metrisch skalierten Merkmalen. Für den Fall, dass Datenpaare mehrfach vorkommen, ist sie definiert als:

$$S_{XY} = \frac{1}{n-1} n_j (x_i - \bar{x})(y_j - \bar{y})$$

- $S_{XY} = \frac{1}{20-1} \{3 \cdot (1-2) \cdot (0-1) + 2 \cdot (1-2) \cdot (1-1) + 1 \cdot (1-2) \cdot (2-1) + \ldots + 3 \cdot (3-2) \cdot (2-1)\} = \frac{3}{19} \approx 0{,}1579$
- $r = \frac{S_{XY}}{S_X S_Y} = \frac{0{,}1579}{0{,}79 \cdot 0{,}92} \approx 0{,}22$
- Es liegt ein schwacher positive linearer Zusammenhang zwischen der Anzahl der Zimmer und der Anzahl der Computer pro Haushalte vor.

Aufgabe 3.2 der AS

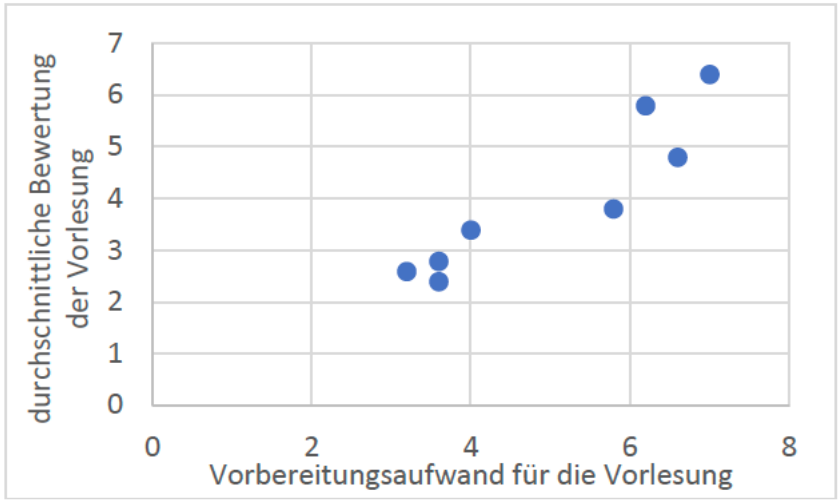
Für acht Termine einer Vorlesung stehen folgende Daten für den Borbereitungsaufwand für die jeweilige Vorlesungen Arbeitstage (X) und für die durchschnittliche Bewertung der jeweiligen Vorlesung auf einer Skala von 1 (sehr schlecht) und 7 (sehr gut) durch Studierende (Y) zur Verfügung:

Vorbereitungsaufwand für die Vorlesung in Arbeitstagen (Merkmal X)	Durchschnittliche Bewertung der Vorlesung (Merkmal Y)
4,0	3,4
3,2	2,6
5,8	3,8
3,6	2,8
6,6	4,8
6,2	5,8
3,6	2,4
7,0	6,4

a) Zeichnen Sie ein Streudiagramm.
b) Berechnen und interpretieren Sie den Korrelationskoeffizienten.
c) Durch eine lineare Regression soll die durchschnittliche Bewertung einer Vorlesung durch den Vorlesungsaufwand erklärt werden. Ermitteln Sie die Gleichung der Regressionsgeraden $\hat{y} = a + bx$ nach der Methode der kleinsten Quadrate und zeichnen Sie die Regressionsgerade in das Streudiagramm ein.

Aufgabe 3.2 der AS - Lösung (1/3)

a)



Aufgabe 3.2 der AS - Lösung (2/3)

b)

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}(4{,}0 + 3{,}2 + 5{,}8 + 3{,}6 + 6{,}6 + 6{,}2 + 3{,}6 + 7{,}0) = 5$
- $S_X^2 = \frac{1}{n\text{-}1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{8\text{-}1}\{(4-5)^2 + (3{,}2-5)^2 + (5{,}8-5)^2 + \ldots + (7{,}0-5)^2\} = 2{,}4$
- $S_x = \sqrt{S_X^2} = \sqrt{2{,}4} \approx 1{,}5492$

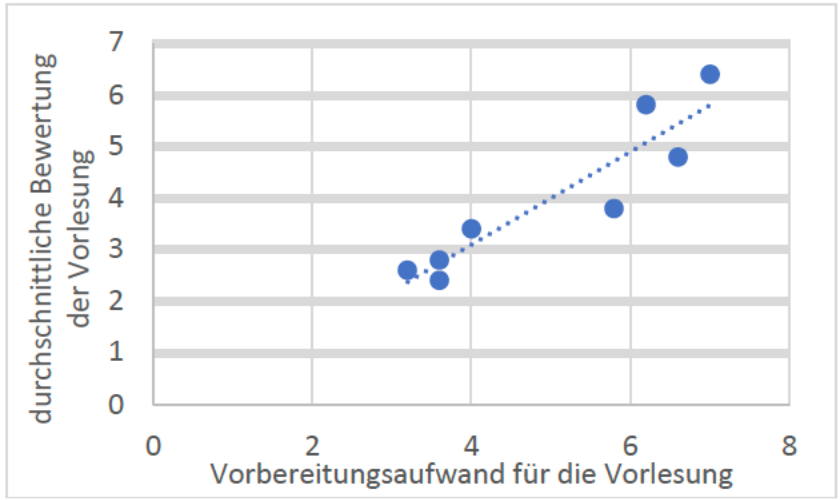
- $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{1}{8}(3{,}4 + 2{,}6 + 3{,}8 + 2{,}8 + 4{,}8 + 5{,}8 + 2{,}4 + 6{,}4) = 4$
- $S_Y^2 = \frac{1}{n\text{-}1}\sum_{j=1}^{n} (y_j - \bar{y})^2 = \frac{1}{8\text{-}1}\{(3{,}4-4)^2 + (2{,}6-4)^2 + (3{,}8-4)^2 + \ldots + (6{,}4-4)^2\} \approx 2{,}2857$
- $S_Y = \sqrt{S_Y^2} \approx \sqrt{2{,}2857} \approx 1{,}5119$

- $S_{XY} = \frac{1}{n\text{-}1}\sum_{i=1}^{n} (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{8\text{-}1}\{(4-5)\cdot(3{,}4-4) + (3{,}2-5)\cdot(2{,}6-4) + (5{,}8-5)\cdot(3{,}8 - 4) + \ldots + (7{,}0-5)\cdot(6{,}4-4)\} = 2{,}16$
- $r = \frac{S_{XY}}{S_X S_Y} = \frac{2{,}16}{1{,}5492 \cdot 1{,}5119} \approx 0{,}9222$
- Es besteht eine sehr starke positive lineare Abhängigkeit zwischen den beiden Merkmalen

Aufgabe 3.2 der AS - Lösung (3/3)

c)

- $b = \frac{s_{XY}}{s_X^2} = \frac{2{,}16}{2{,}4} = 0{,}9$
- $a = \bar{y} - b\bar{x} = 4 - 0{,}9 \cdot 5 = -0{,}5$
- $\hat{y} = -0{,}5 + 0{,}9x$



it's time for a quiz!