Kapitel 5: Zusammenhangsmaße

Beispiel 5.1

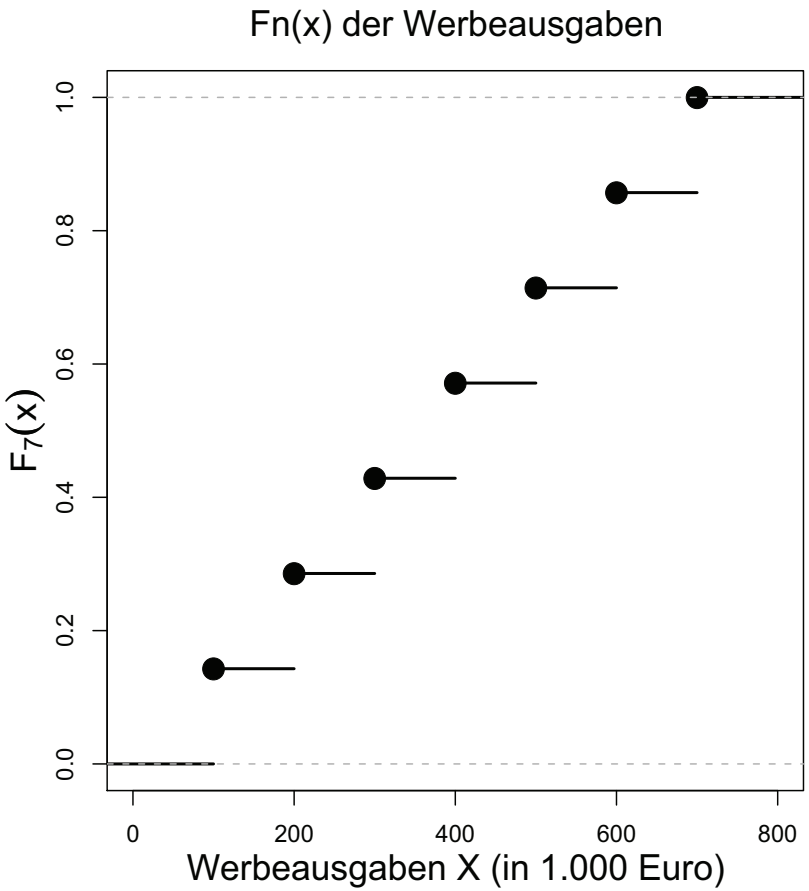
- Werbeausgaben und Umsätze verschiedener Firmen (fiktiv)

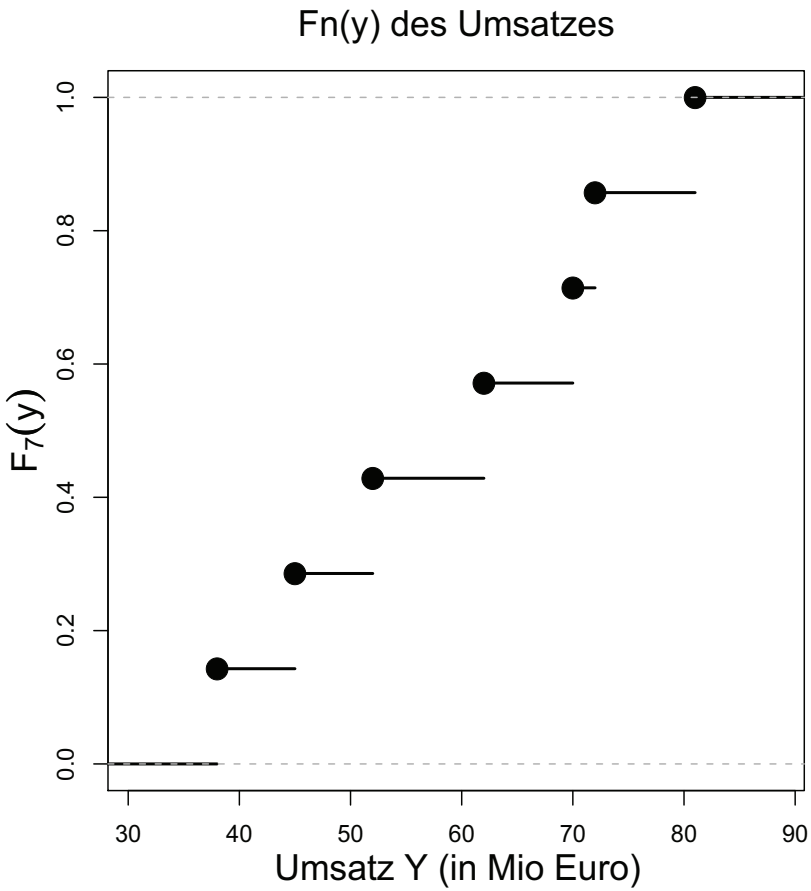
Firma Nr. i	Werbeausgaben X_i (in 1.000 Euro)	Umsatz Y_i (in Mio. Euro)
1	100	38
2	200	45
3	300	52
4	400	62
5	500	72
6	600	70
7	700	81

$\longrightarrow$ Struktur der Daten?

Beispiel 5.1 (Fortsetzung)

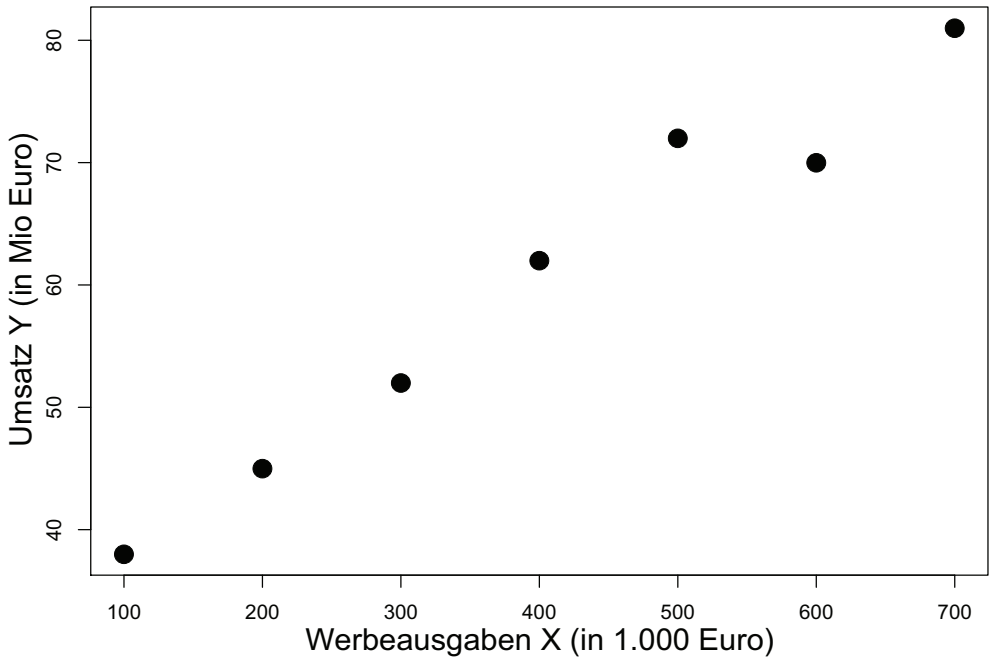
- Kapitel 1-4: Betrachte für Merkmal X und Merkmal Y etwa die empirischen Verteilungsfunktionen





Beispiel 5.1 (Fortsetzung)

- Mittelwert und Varianz der Merkmale X und Y :
  $\bar{x}^a = 400,\ s_x^2 = 40.000; \quad \bar{y}^a = 60,\ s_y^2 = 208,86$
- Trage Ausprägung x_i gegen Ausprägung y_i ab



$\rightarrow$ (positiver) Zusammenhang von X und Y, der weder von emp. Verteilungsfunktion, Mittelwert noch Varianz berücksichtigt wird $\rightarrow$ Zusammenhangsmaß vonnöten

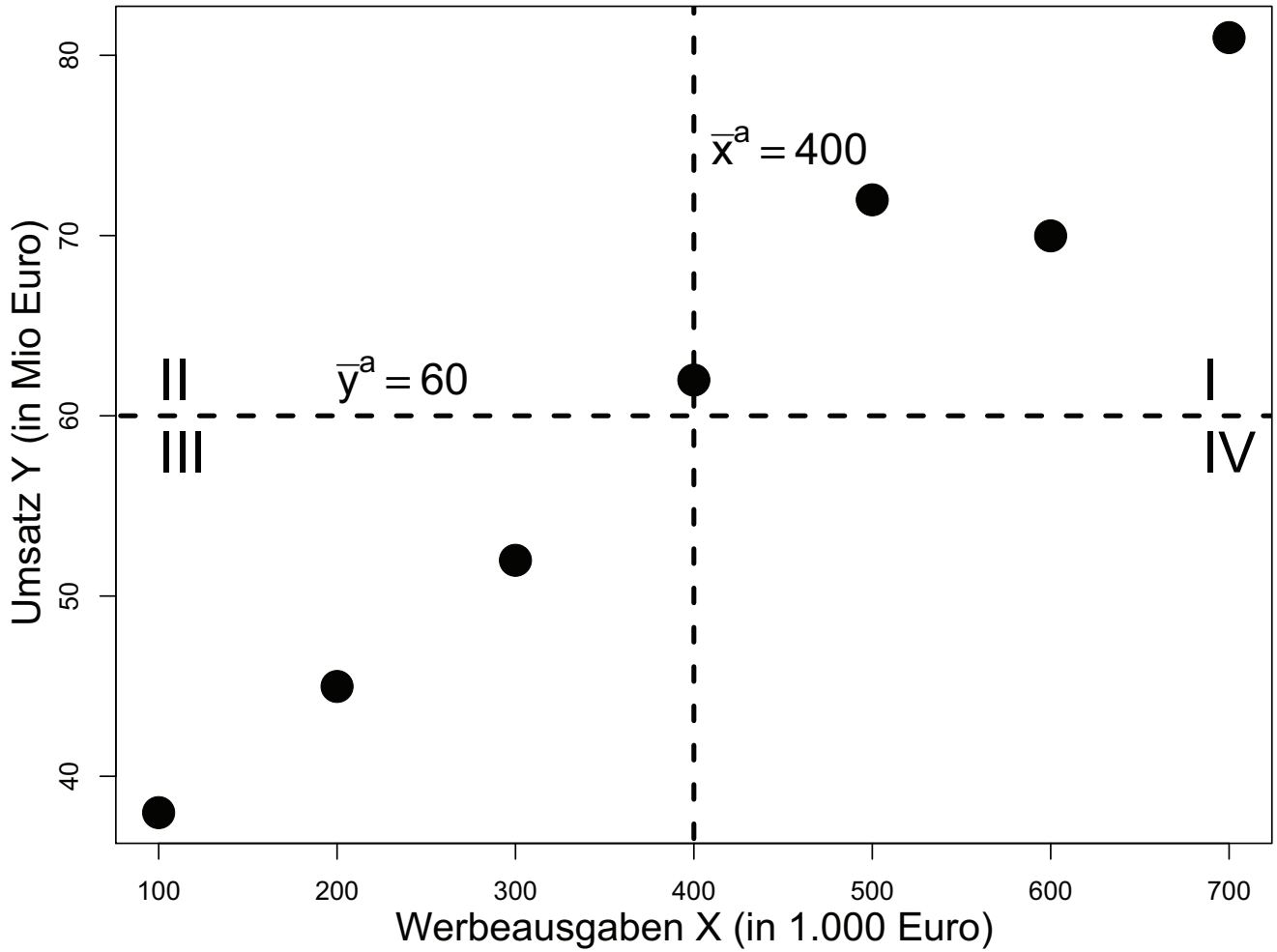
Bemerkung

- Bisher: Ein Merkmal pro Merkmalsträger
- Jetzt: Zwei Merkmale pro Merkmalsträger
- Gesucht: Maßzahlen, die den Zusammenhang zwischen diesen beiden Merkmalen beschreiben

Beispiel 5.2

(Umsätze & Werbeausgaben von Firma i, vgl. Beispiel 5.1)

- Eine Möglichkeit: Einteilung des Koordinatensystems in vier Quadranten durch Mittelwerte



Beispiel 5.2 (Fortsetzung)

- Idee nun
  - Häufung der Beobachtungen in den Quadranten I und III $\rightarrow$ positiver Zusammenhang
  - Häufung der Beobachtungen in den Quadranten II und IV $\rightarrow$ negativer Zusammenhang
  - Ähnlich große Beobachtungszahlen in den Quadrantenpaaren (I,III) und (II,IV) $\rightarrow$ kein Zusammenhang

- Hier:

$$\begin{array}{lllllllll} \text{I} & + & \text{III} & = & 3{,}5 & + & 3 & = & 6{,}5 \\ \text{II} & + & \text{IV} & = & 0{,}5 & + & 0 & = & 0{,}5 \end{array}$$

$\rightarrow$ „stark“ positiver Zusammenhang (fasse hierbei die Beobachtung $(x_4, y_4) = (400, 62)$ als halb zum ersten und halb zum zweiten Quadranten zugehörig auf)

Bemerkung

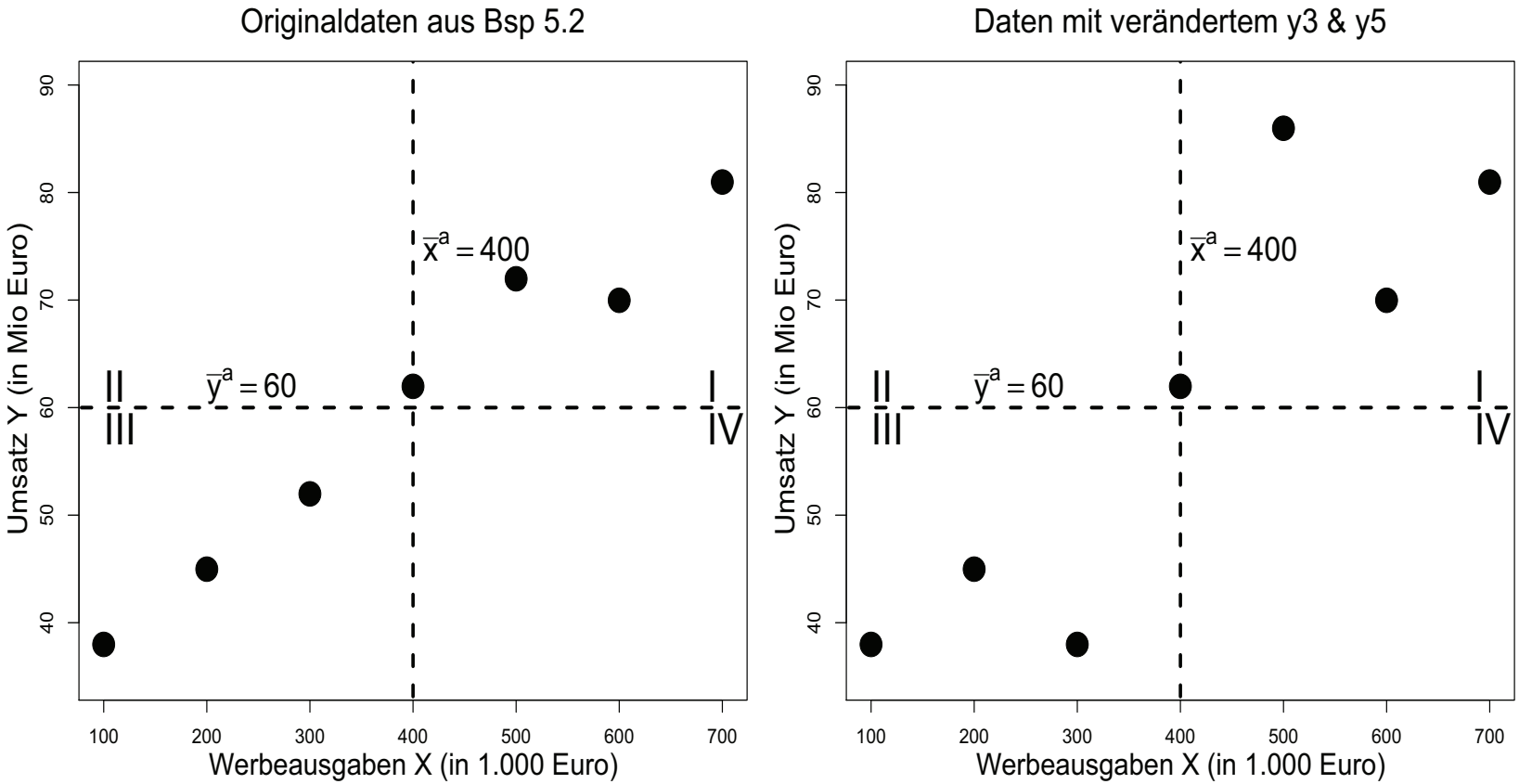
a) Kriterium aus Beispiel 5.2 recht grob, Entfernung der Beobachtungen vom „Zentrum“ $(\bar{x}^a, \bar{y}^a)$ wird nicht berücksichtigt $\rightarrow$

- Betrachte abermals Umsätze und Werbeausgaben aus Beispiel 5.1 und 5.2
- Beobachtung $y_3 = 52$ Mio. Euro verändere sich zu $y_3^{neu} = 38$ Mio. Euro
- Beobachtung $y_5 = 72$ Mio. Euro verändere sich zu $y_5^{neu} = 86$ Mio. Euro

$\rightarrow$ $\bar{y}^a_{neu} = \bar{y}^a = 60$ Mio. Euro ($\bar{x}^a_{neu} = \bar{x}^a = 400.000$ Euro sowieso)

Bemerkung (Fortsetzung)

a) (Fortsetzung)



$\longrightarrow$ Gemäß des Kriteriums aus Beispiel 5.2 ist es egal, ob sich Beobachtungen y_3 und y_5 oder y_3^{neu} und y_5^{neu} realisieren, der Zusammenhang bleibt gleich stark

Bemerkung (Fortsetzung)

b) Motiviert durch Teil a): Fordere unterschiedliche Gewichtung der Daten, je nach Entfernung von $(\bar{x}^a, \bar{y}^a)$ $\rightarrow$ Gewicht für Beobachtungspaar i : $(x_i - \bar{x}^a)(y_i - \bar{y}^a)$

- $x_i > \bar{x}^a$ und $y_i > \bar{y}^a \Rightarrow (x_i - \bar{x}^a)(y_i - \bar{y}^a) > 0$ (Quadr. I)
- $x_i < \bar{x}^a$ und $y_i < \bar{y}^a \Rightarrow (x_i - \bar{x}^a)(y_i - \bar{y}^a) > 0$ (Quadr. III)
- $x_i < \bar{x}^a$ und $y_i > \bar{y}^a \Rightarrow (x_i - \bar{x}^a)(y_i - \bar{y}^a) < 0$ (Quadr. II)
- $x_i > \bar{x}^a$ und $y_i < \bar{y}^a \Rightarrow (x_i - \bar{x}^a)(y_i - \bar{y}^a) < 0$ (Quadr. IV)
- $x_i = \bar{x}^a$ oder $y_i = \bar{y}^a \Rightarrow (x_i - \bar{x}^a)(y_i - \bar{y}^a) = 0$

$\rightarrow$ Berechne $(x_i - \bar{x}^a)(y_i - \bar{y}^a)$ für alle Beobachtungspaare und betrachte den Durchschnitt

Definition 5.1

Für zwei kardinal skalierte Merkmale X und Y mit den beobachteten Ausprägungen $x_1, x_2, \ldots, x_n$ und $y_1, y_2, \ldots, y_n$ heißt

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

Kovarianz (oder gemeinsame Streuung) von X und Y.

Beispiel 5.3

(Umsätze & Werbeausgaben von Firma i, vgl. Beispiele 5.1 und 5.2)

x_i	y_i	$x_i - \bar{x}^a$	$y_i - \bar{y}^a$	$(x_i - \bar{x}^a) \cdot (y_i - \bar{y}^a)$
100	38	-300	-22	6.600
200	45	-200	-15	3.000
300	52	-100	-8	800
400	62	0	2	0
500	72	100	12	1.200
600	70	200	10	2.000
700	81	300	21	6.300
$\sum$ 2.800	420	0	0	19.900

$\rightarrow \quad s_{xy} = 1/7 \times 19.900 = 2.842,86$

Beispiel 5.3 (Fortsetzung)

- Für die veränderten Daten aus Bemerkung b) nach Beispiel 5.2 ($y_3 \to y_3^{neu},\ y_5 \to y_5^{neu}$) ergibt sich $s_{xy}^{neu} = 3242,86$

Bemerkung

a) Für die Kovarianz gilt $s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \bar{x}^a \cdot \bar{y}^a$

b) Betrachte lineare Transformationen der Form

$x_i^\star = a \cdot x_i + b$ und $y_i^\star = c \cdot y_i + d \quad (a, b, c, d \in \mathbb{R},\ i = 1, ..., n)$,

dann gilt $s_{x^\star y^\star} = a \cdot c \cdot s_{xy}$

$\to$ Kovarianz ist abhängig von der Maßeinheit

c) s_{xy} repräsentiert Richtung des Zusammenhangs zwischen zwei Variablen (positiv $\to s_{xy} > 0$, negativ $\to s_{xy} < 0$); keine Aussage über Stärke des Zusammenhangs möglich

Beispiel 5.4

(Umsätze & Werbung, vgl. Beispiele 5.1 bis 5.3)

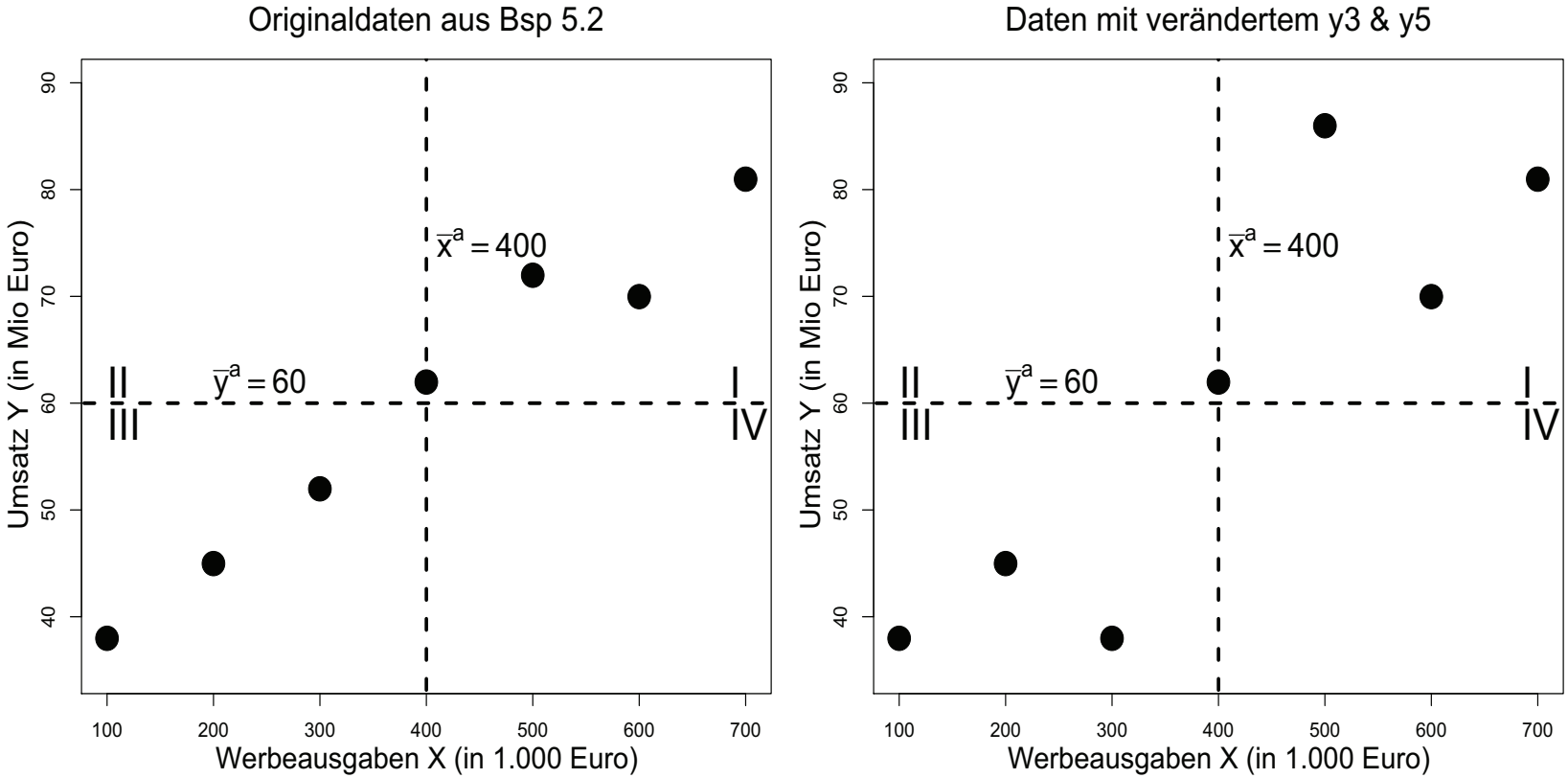
a) Messe Werbeausgaben nun in $100 (= 1.000/10)$ Euro, Umsatz in $100.000 (= 1.000.000/10)$ Euro

x_i	y_i	$x_i - \bar{x}^a$	$y_i - \bar{y}^a$	$(x_i - \bar{x}^a) \cdot (y_i - \bar{y}^a)$
1.000	380	-3.000	-220	660.000
2.000	450	-2.000	-150	300.000
3.000	520	-1.000	-80	80.000
4.000	620	0	20	0
5.000	720	1.000	120	120.000
6.000	700	2.000	100	200.000
7.000	810	3.000	210	630.000
$\sum$ 28.000	4.200	0	0	1.990.000

$$\rightarrow \quad s_{xy} = 1/7 \times 1.990.000 = 284285,7 = 10 \times 10 \times 2.842,86$$

Beispiel 5.4 (Fortsetzung)

b) Betrachte (neben Daten aus Bsp. 5.1 und 5.2) noch einmal die veränderten Ausprägungen aus Bem. a) nach Bsp. 5.2



$\longrightarrow$ Grafik: Positiver Zusammenhang bei Originaldaten stärker; dies durch Kovarianzen nicht quantifiziert ($s_{xy} = 2.842,86$ und $s_{xy}^{neu} = 3242,86$), vgl. Bem. c) nach Bsp. 5.3

Definition 5.2

Für zwei kardinal skalierte Merkmale X und Y mit den beobachteten Ausprägungen $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ heißt

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}^a\right)\left(y_i - \bar{y}^a\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}^a\right)^2 \cdot \sum_{i=1}^{n} \left(y_i - \bar{y}^a\right)^2}}$$

Bravais-Pearson-Korrelationskoeffizient von X und Y.

Beispiel 5.5

(Umsätze & Werbung, vgl. Beispiele 5.1 bis 5.4)

a) Für die Originaldaten aus den Beispielen 5.1 und 5.2 ergibt sich

$$\begin{aligned} s_x^2 &= 40.000 \quad \text{und} \quad s_y^2 = 208,86 \\ r_{xy} &= \frac{2842,86}{\sqrt{40.000 \cdot 208,86}} = 0,984 \end{aligned}$$

$\rightarrow$ Umrechnung der Maßeinheiten in 100 Euro (Werbung) bzw. 100.000 Euro (Umsatz) verändert diesen Wert nicht

$$r_{x^\star y^\star} = \frac{284285,7}{\sqrt{4.000.000 \cdot 20886}} = 0,984$$

Beispiel 5.5 (Fortsetzung)

b) Datenvariation aus Bemerkung a) nach Beispiel 5.2

$$s_x^{2,neu} = 40.000 \quad \text{und} \quad s_y^{2,neu} = 344,86$$

$$r_{xy}^{neu} = \frac{3242,86}{\sqrt{40.000 \cdot 344,86}} = 0,873 < 0,984 = r_{xy}$$

$\rightarrow$ Zusammenhang der veränderten Daten „schwächer“

Bemerkung

(Eigenschaften von $s_{xy},\ r_{xy}$)

a) Für die Kovarianz gilt

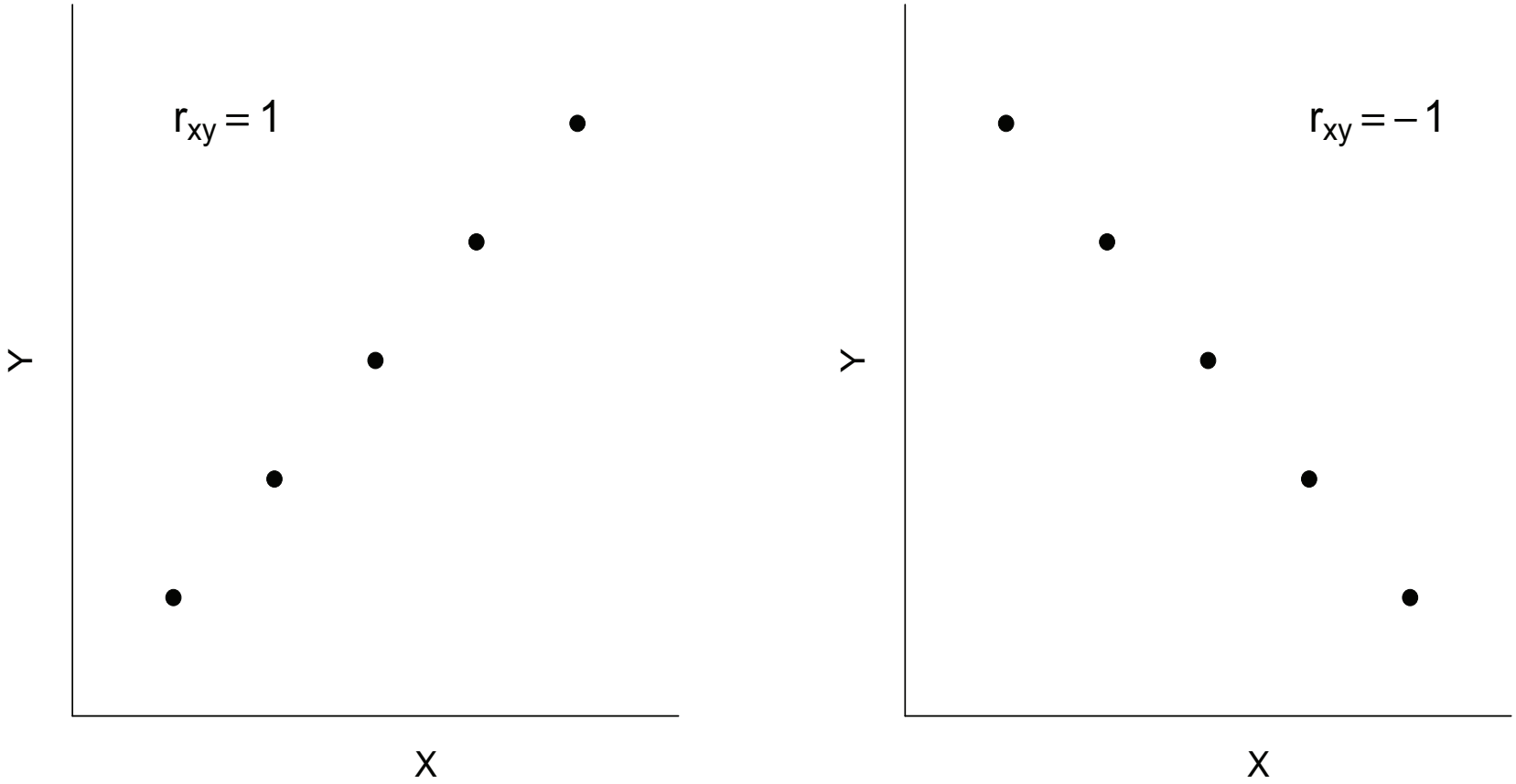
$$\begin{aligned} &\text{(I)} && |s_{xy}| && \leq && s_x \cdot s_y \\ &\text{(II)} && |s_{xy}| && = && s_x \cdot s_y \quad \Leftrightarrow \ y_i = a \cdot x_i + b \text{ mit } a \neq 0, \end{aligned}$$

also gilt für den Bravais-Pearson-Korrelationskoeffizienten

$$\begin{aligned} &\text{(III)} && -1 \leq r_{xy} \leq 1 \\ &\text{(IV)} && r_{xy} = 1 \ \Leftrightarrow \ y_i = a \cdot x_i + b \text{ mit } a > 0 \\ &\text{(V)} && r_{xy} = -1 \ \Leftrightarrow \ y_i = a \cdot x_i + b \text{ mit } a < 0 \end{aligned}$$

Bemerkung (Fortsetzung)

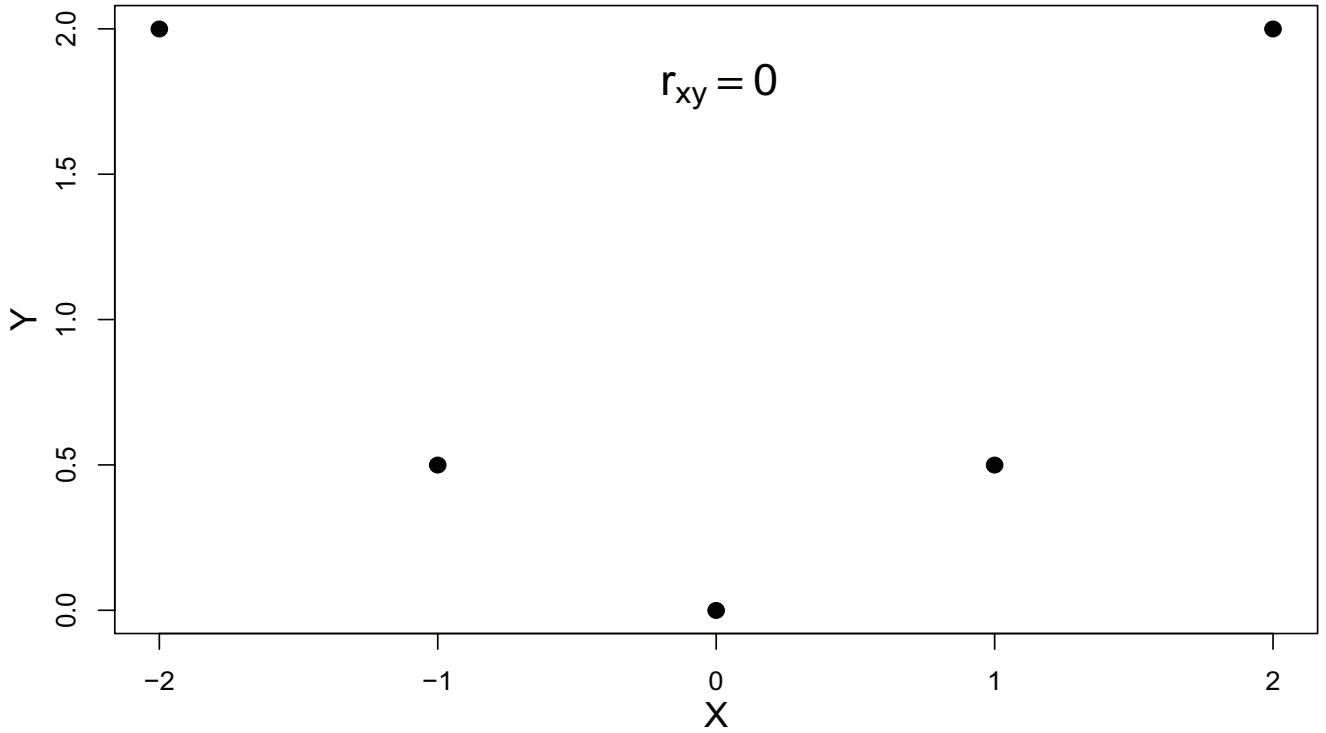
b) Bedeutung von Bemerkung a), (IV) & (V)



Bemerkung (Fortsetzung)

c) Vorsicht: $r_{xy} = 0$ heißt nicht, dass kein Zusammenhang besteht, sondern dass kein linearer Zusammenhang vorliegt;

- Betrachte Merkmal X mit Ausprägungen $-2, -1, 0, 1, 2$ und Merkmal Y mit Ausprägungen $y_i = 0.5 \times x_i^2$ (d.h. Merkmal X erklärt Merkmal Y komplett) $\rightarrow$ $r_{xy} = 0$!



Bemerkung (Fortsetzung)

d) Vorsicht: Korrelation ist nicht gleich Kausalität, Zusammenhang kann etwa durch dritte Einflussgröße Verursacht werden

- Beispiel 1 (aus www.statistics4u.info): Schuhgröße und Kalziumgehalt der Knochen positiv korreliert; Grund: Kinder haben weniger Kalzium in den Knochen als Erwachsene, und natürlich geringere Schuhgrößen
- Beispiel 2: Zahl der Störche und Kinderanzahl pro Ehepaar positiv korreliert; Grund: Je ländlicher die Gegend, umso mehr Störche gibt es, und umso mehr Kinder werden pro Ehepaar geboren

$\longrightarrow$ halte den dritten Faktor (in Bsp. 1 etwa das Alter und in Bsp. 2 die Größe der untersuchten Stadt) konstant
$\longrightarrow$ beide „Korrelationen“ verschwinden

Beispiel 5.6

- Erhebe an 11 Studenten die Punktezahlen in der Statistik- bzw. Mathematik-Klausur (vgl. Bamberg et al., 2007)

Student	A	B	C	D	E	F	G	H	I	J	K
Mathe	38	47	44	51	35	29	22	14	12	19	9
Statistik	39	34	31	48	46	23	17	12	16	28	10

$\longrightarrow$ Zusammenhang der Merkmale?

- Problem bei Bravais-Pearson-Koeffizient: Kardinales Skalenniveau hier zumindest fragwürdig
  - Annahme: Ab 20 Punkten ist die Mathematikklausur bestanden $\longrightarrow$ Abstand zwischen 19 und 20 Punkten sicherlich größer, als etwa zwischen 35 und 36 Punkten
  - Umskalierungen bei Punktevergabe möglich

Definition 5.3

Betrachte zwei Merkmale X und Y mit mindestens ordinalem Skalenniveau und Ausprägungen $x_1, ..., x_n$ bzw. $y_1, ..., y_n$. Die Beobachtung x_k stehe in der Reihe $x_{(1)}, ..., x_{(n)}$ der aufsteigend geordneten Daten an Stelle l (d.h. $x_k = x_{(l)}$). Dann heißt

$$R(x_k) = l$$

Rang von x_k ($R(y_i)$ analog) und

$$r_{xy}^R = \frac{\sum_{i=1}^{n} \left(R(x_i) - \bar{R}_x^a\right)\left(R(y_i) - \bar{R}_y^a\right)}{\sqrt{\sum_{i=1}^{n} \left(R(x_i) - \bar{R}_x^a\right)^2 \cdot \sum_{i=1}^{n} \left(R(y_i) - \bar{R}_y^a\right)^2}}$$

Rangkorrelationskoeffizient nach Spearman

Beispiel 5.7

(Klausurpunkte Mathe & Statistik, vgl. Bsp. 5.6)

Student	A	B	C	D	E	F	G	H	I	J	K
x_i	38	47	44	51	35	29	22	14	12	19	9
$R(x_i)$	8	10	9	11	7	6	5	3	2	4	1
y_i	39	34	31	48	46	23	17	12	16	28	10
$r(y_i)$	9	8	7	11	10	5	4	2	3	6	1

$$\rightarrow \quad \bar{R}_x^a = \bar{R}_y^a = 6$$

Beispiel 5.7 (Fortsetzung)

Stud.	$R(x_i) - \bar{R}_x^a = M_i$	M_i^2	$R(y_i) - \bar{R}_y^a = S_i$	S_i^2	$M_i \cdot S_i$
A	2	4	3	9	6
B	4	16	2	4	8
C	3	9	1	1	3
D	5	25	5	25	25
E	1	1	4	16	4
F	0	0	-1	1	0
G	-1	1	-2	4	2
H	-3	9	-4	16	12
I	-4	16	-3	9	12
J	-2	4	0	0	0
K	-5	25	-5	25	25
$\sum$	0	110	0	110	97

$$\rightarrow \quad r_{xy}^R = \frac{97}{\sqrt{110 \cdot 110}} = 0,88$$

Bemerkung

(Eigenschaften des Rangkorrelationskoeffizienten nach Spearman)

a) $-1 \leq r_{xy}^R \leq 1$

b) $r_{xy}^R = 1 \Leftrightarrow R(x_i) = R(y_i)$ für alle i

c) $r_{xy}^R = -1 \Leftrightarrow R(x_i) = n - R(y_i) + 1$ für alle i

Bemerkung (Fortsetzung)

d) Gemäß Teil b) und c) misst r_{xy}^R den monotonen Zusammenhang zweier Merkmale (im Gegensatz zum Bravais-Pearson-Koeffizienten, der den linearen Zusammenhang misst)