Bsp. Bier und Kopfweh

- Fiktive Kontingenztafel zum Bierkonsum auf einer Party und Kopfschmerzen am nächsten Tag:

Bier \ Kopfweh	ja	nein	
ja	16	40	56
nein	7	28	35
	23	68	91

- Gibt es einen statistischen Zusammenhang zwischen den Merkmalen „Bier“ und „Kopfweh“?

Empirische Unabhängigkeit

- Es besteht keinerlei Zusammenhang zwischen den Merkmalen X und Y.
- Frage: Wie sollten die Häufigkeiten verteilt sein, wenn die beiden Merkmale keinerlei Zusammenhang aufweisen?

Merkmal X \ Merkmal Y	b_1	b_2	
a_1	?	?	$h_{1\cdot}$
a_2	?	?	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	n

Postulat der emp. Unabhängigkeit

- Bezeichnet $\tilde{h}_{ij}$ die Häufigkeit, die man erwarten würde, wenn kein Zusammenhang vorliegt, dann gilt

$$\frac{\tilde{h}_{ij}}{h_{i\cdot}} = \frac{h_{\cdot j}}{n} \quad \Longleftrightarrow \quad \tilde{h}_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}$$

bzw.

$$\frac{\tilde{h}_{ij}}{h_{\cdot j}} = \frac{h_{i\cdot}}{n} \quad \Longleftrightarrow \quad \tilde{h}_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}\,.$$

- Idee: Nutze die Diskrepanz zwischen den beobachteten und erwarteten Werten zur Konstruktion eines Zusammenhangsmaßes.

χ^2-Koeffizient

- Der χ^2**-Koeffizient** misst, wie weit die beobachtete Kontingenztabelle von einer Tabelle entfernt ist, die das Postulat der empirischen Unabhängigkeit erfüllt.
- Seien h_{ij} die tatsächlich beobachteten Häufigkeiten.
- Seien $\tilde{h}_{ij} = h_{i\cdot}h_{\cdot j}/n$ die Häufigkeiten, die zu erwarten sind, wenn kein Zusammenhang vorliegt.
- Dann ist der χ^2-Koeffizient bestimmt durch

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(h_{ij} - \tilde{h}_{ij}\right)^2}{\tilde{h}_{ij}}\,, \quad \chi^2 \in [0, \infty)\,.$$

Eigenschaften χ^2-Koeffizient

- χ^2 ist groß, wenn X und Y stark voneinander abhängen (starke Diskrepanz zwischen beobachteten und erwarteten Werten).
- χ^2 ist klein, wenn X und Y schwach voneinander abhängen (kleine Diskrepanz zwischen beobachteten und erwarteten Werten).
- χ^2 ist Null, wenn X und Y nicht voneinander abhängen.
- χ^2 hängt jedoch auch vom Stichprobenumfang n und von der Dimension der Tafel ab.

Kontingenzkoeffizienten

- Der **Kontingenzkoeffizient** ist bestimmt durch

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

und besitzt den Wertebereich $K \in \left[0, \sqrt{\frac{M-1}{M}}\right]$, wobei $M = \min\{k, m\}$ ($k =$ Anzahl Zeilen, $m =$ Anzahl Spalten) ist.

- Der **normierte Kontingenzkoeffizient** ergibt sich durch

$$K^* = K / \sqrt{\frac{M-1}{M}}$$

mit dem Wertebereich $K^* \in [0, 1]$.

Bem. Kontingenzkoeffizienten

- Die Maße χ^2, K, K^* besitzen folgende Eigenschaften:
  - Es wird nur die *Stärke* des Zusammenhangs gemessen, nicht ob ein „positiver“ oder „negativer“ Zusammenhang vorliegt.
  - Die Maße sind vergleichender Art.
  - Sämtliche Maße benutzen nur das Nominalskalenniveau von X und Y.

Visualisierungen von Kontingenztafeln

- zweidimensionales Säulendiagramm
- Rechteckdiagramm
- Mosaik-Plot

Bsp. Zweidim. Säulendiagramm

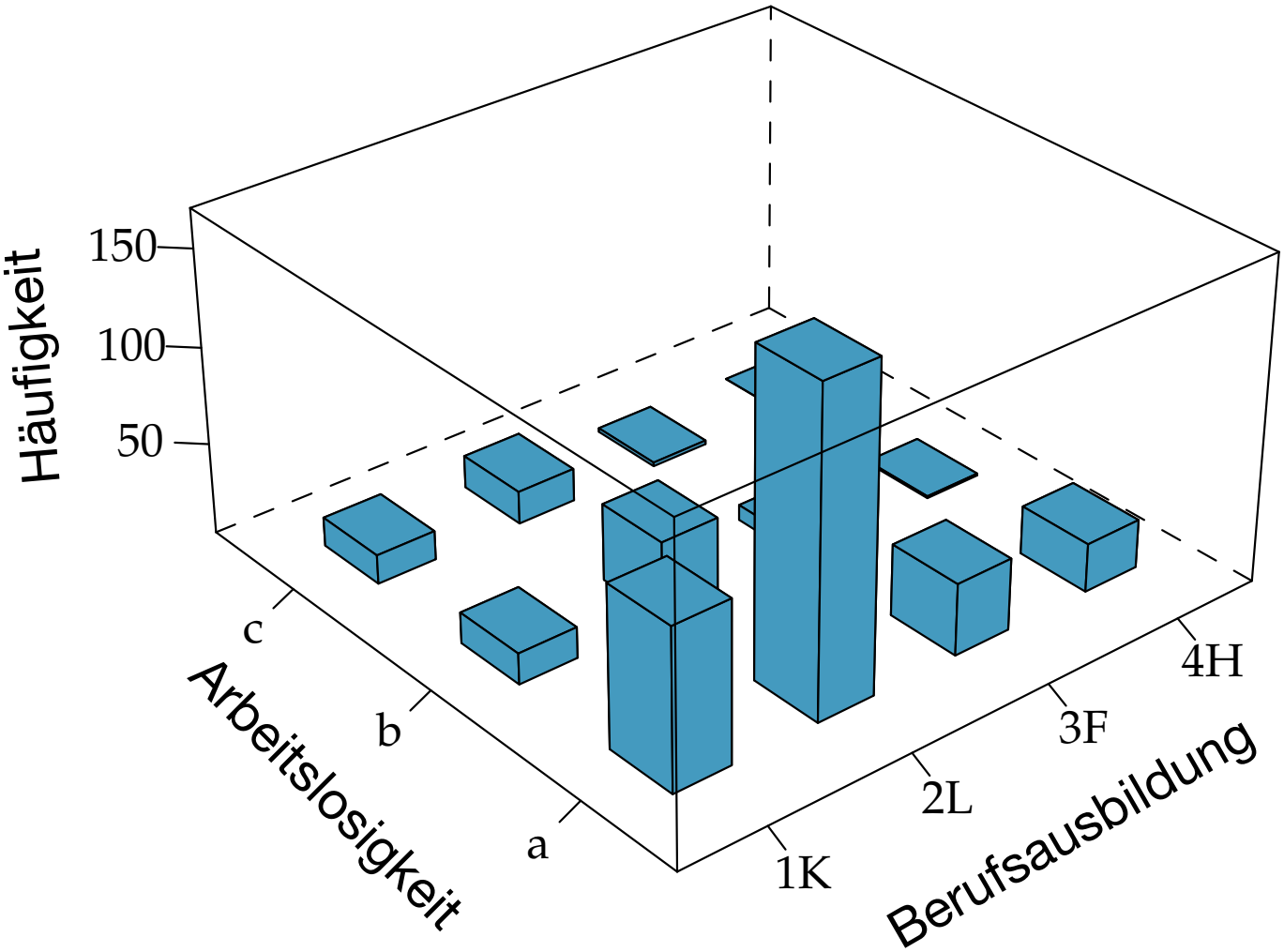


Abb.: Säulendiagramm zur Ausbildung

($1K$: keine Ausbildung, $2L$: Lehre, $3F$: fachspez. Ausbildung, $4H$: Hochschule) und Dauer der Arbeitslosigkeit (a: ≤ 6 Monate, b: 6-12 Monate, c: > 12 Monate)

Bsp. Rechteckdiagramm (1)

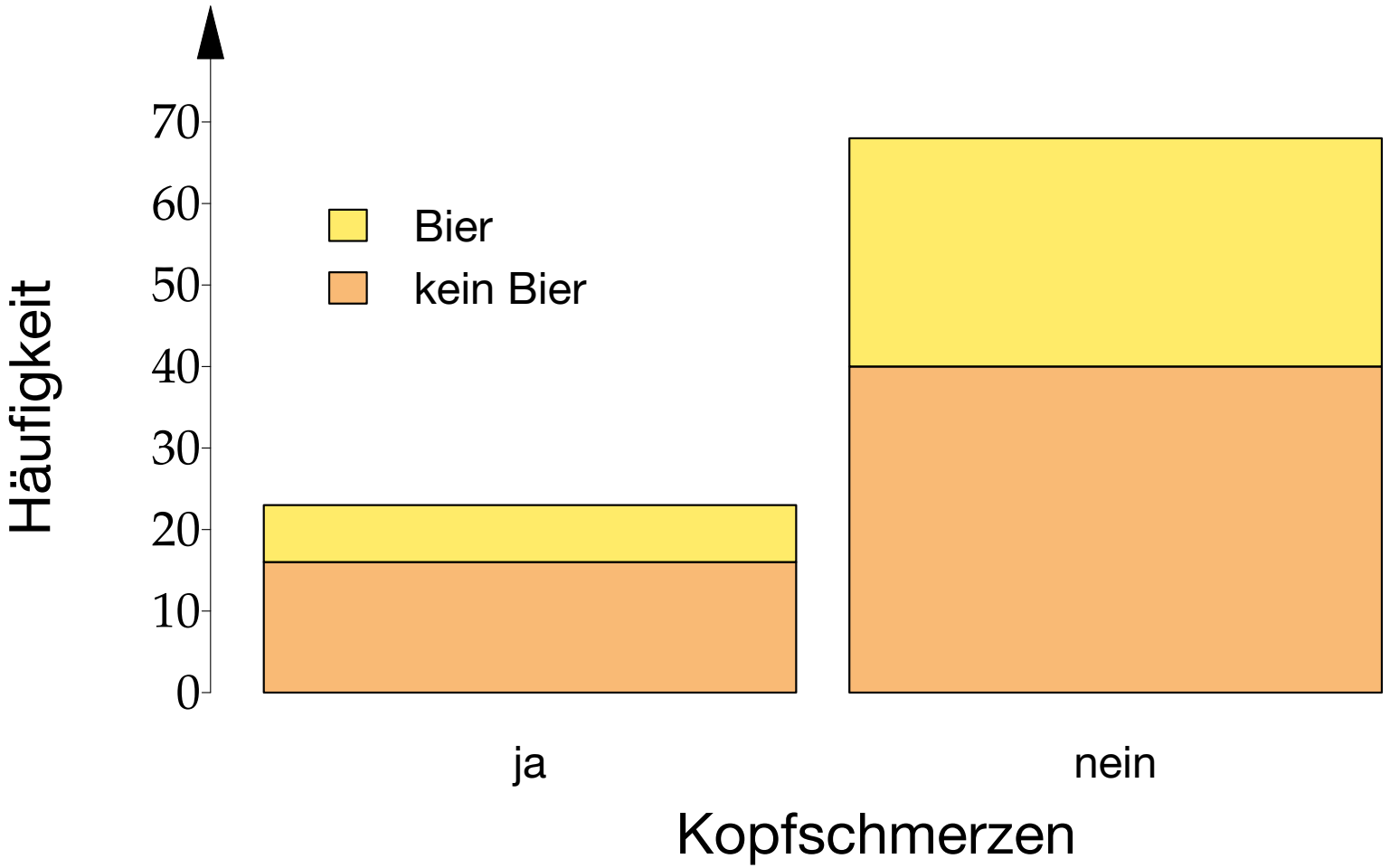


Abb.: Rechteckdiagramm von Bierkonsum und Kopfschmerzen für $n = 91$ Personen

Bsp. Rechteckdiagramm (2)

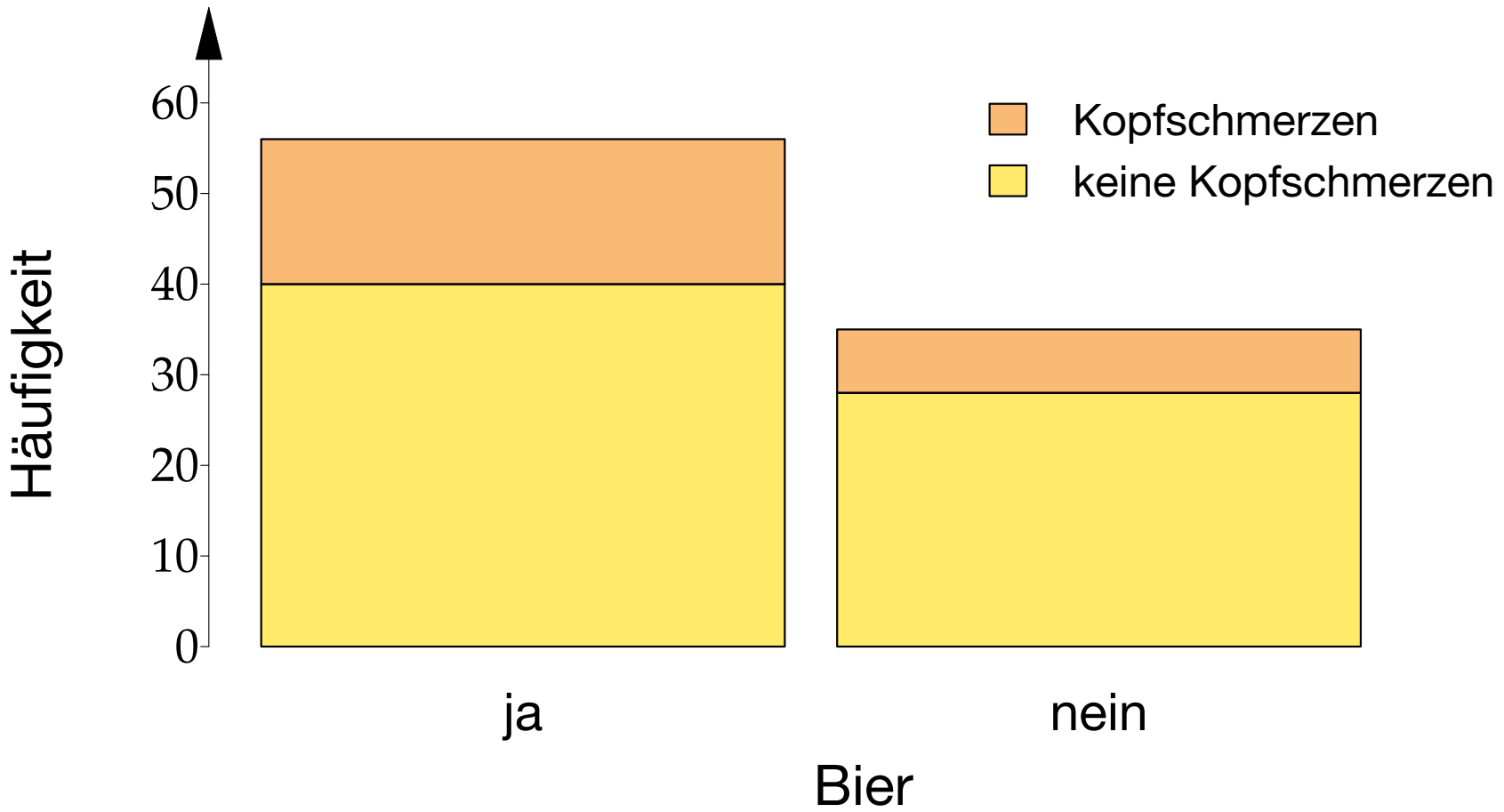


Abb.: Rechteckdiagramm von Bierkonsum und Kopfschmerzen für $n = 91$ Personen

Mosaik-Plot

- flächenproportionale Darstellung von Häufigkeiten der zu Grunde liegenden Kontingenztafel
- Aufteilung der Rechteckflächen mit Bezug auf die entsprechenden Zeilen- bzw. Spaltensummen
- gut geeignet für mehrkategoriale Daten

Bsp. Mosaik-Plot Bier und Kopfweh

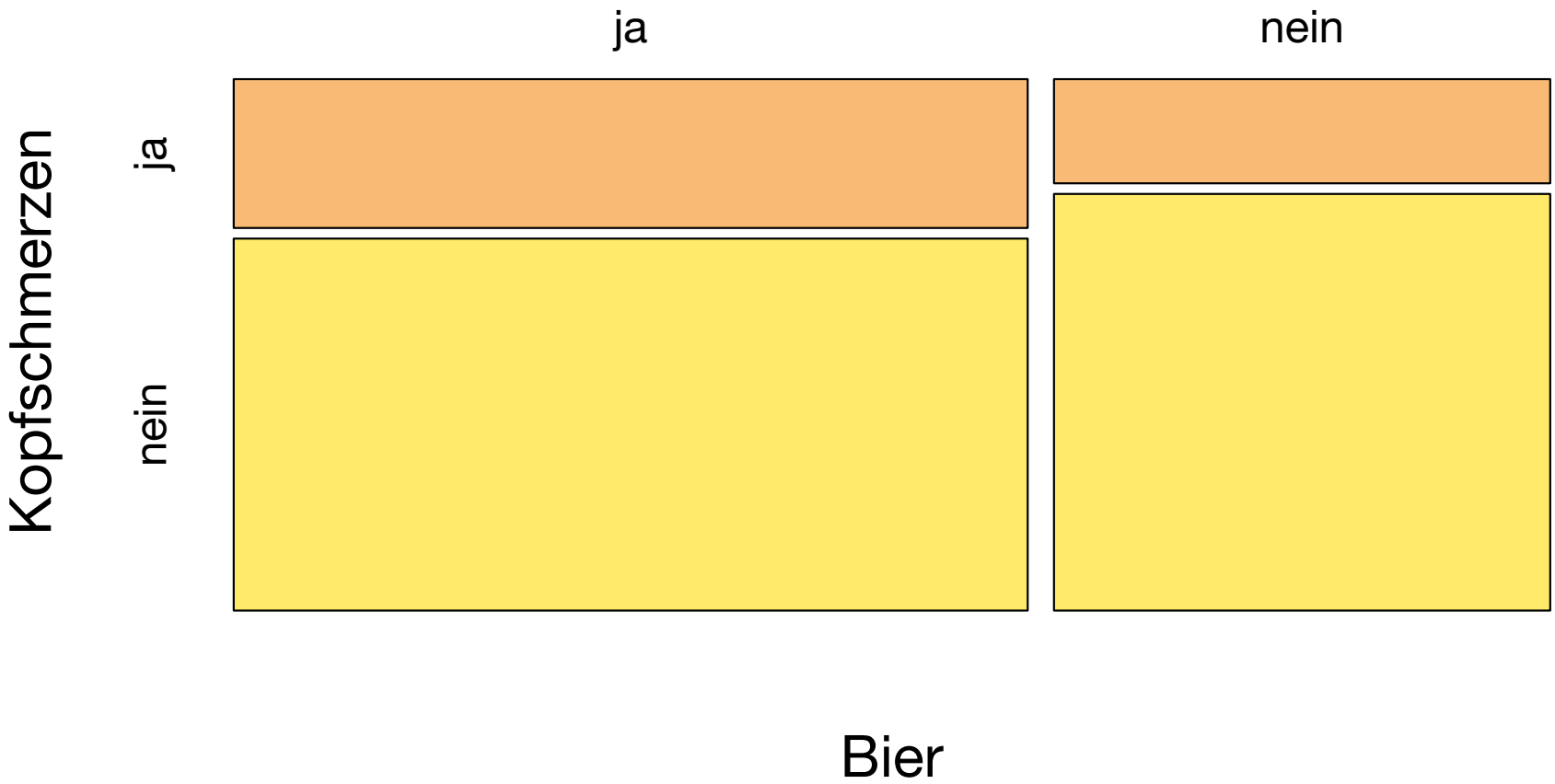


Abb.: Mosaik-Plot von Bierkonsum und Kopfschmerzen für $n = 91$ Personen

Bsp. Daten Titanic-Untergang

- Datensatz (aus R): Überleben beim Titanic-Untergang
- mehrere diskrete Merkmale: Geschlecht, Klasse, Kind/Erwachsener, Überleben (ja/nein)
- grafische Darstellung durch Mosaik-Plot

Bsp. Mosaik-Plot Titanic (1)

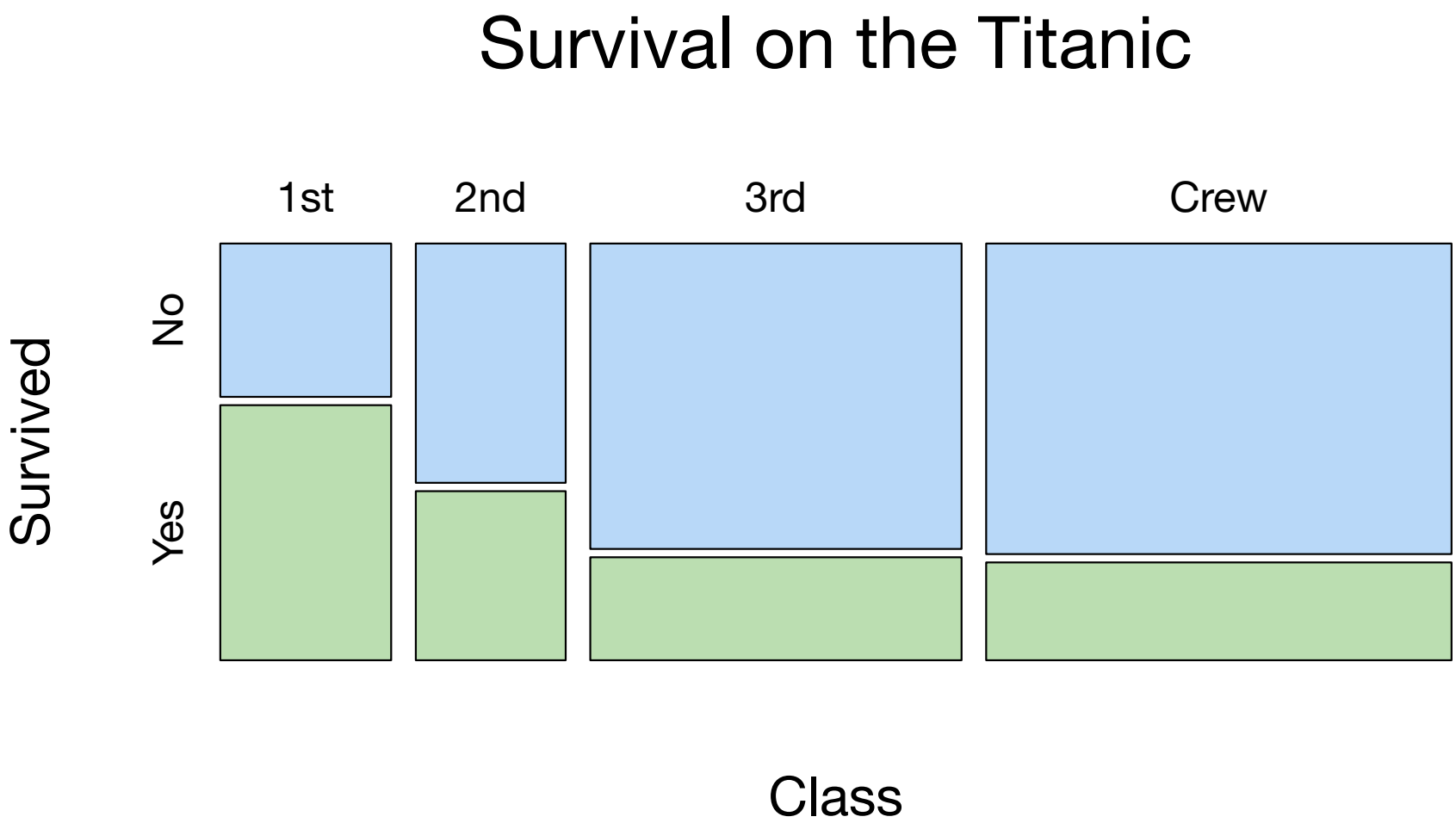


Abb.: Überleben (ja/nein) und Klasse (1/2/3/Besatzung) für $n = 2201$ Passagiere der Titanic

Bsp. Mosaik-Plot Titanic (2)

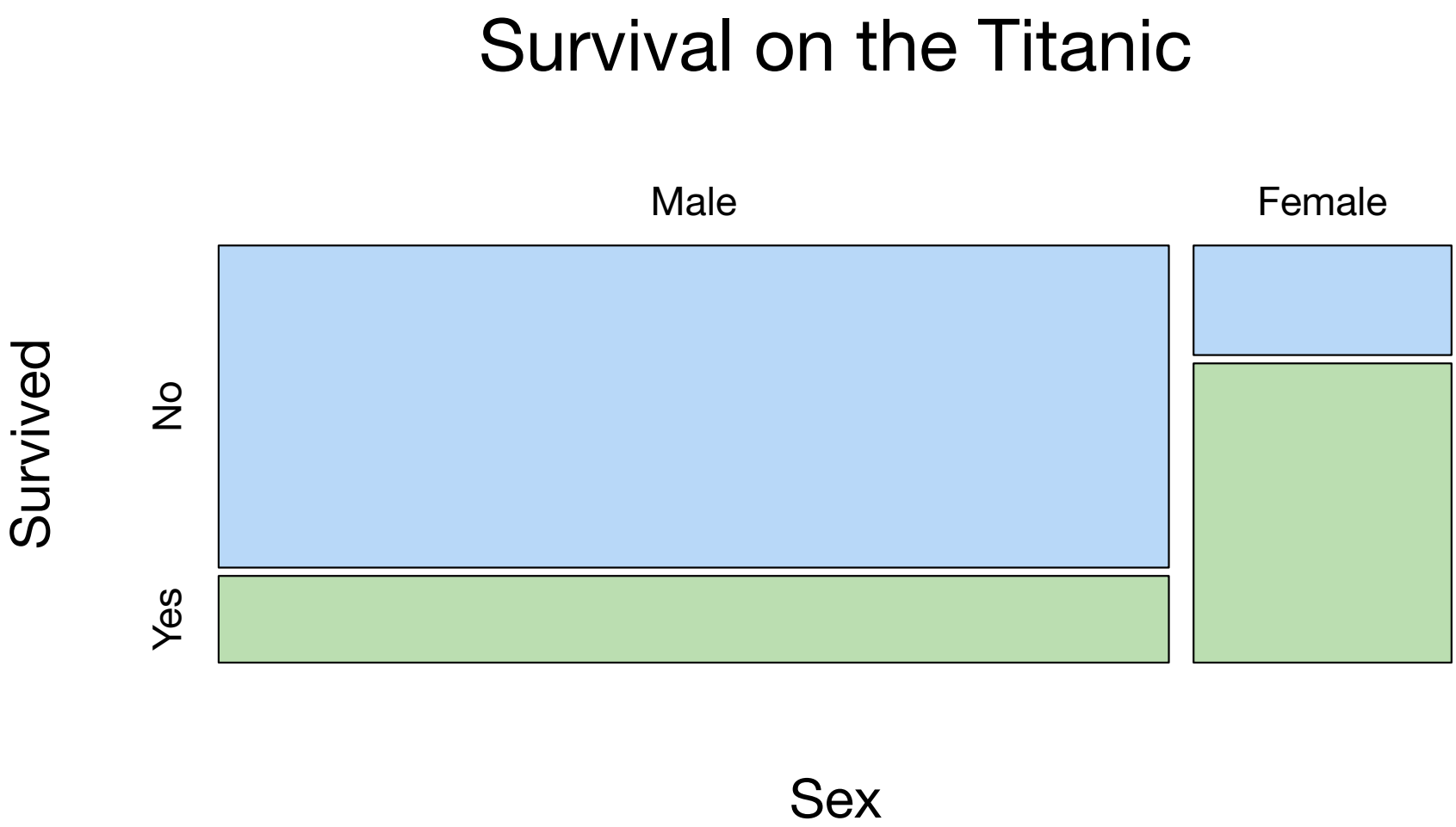


Abb.: Überleben (ja/nein) und Geschlecht (männlich/weiblich) für $n = 2201$ Passagiere der Titanic

Bsp. Mosaik-Plot Titanic (3)

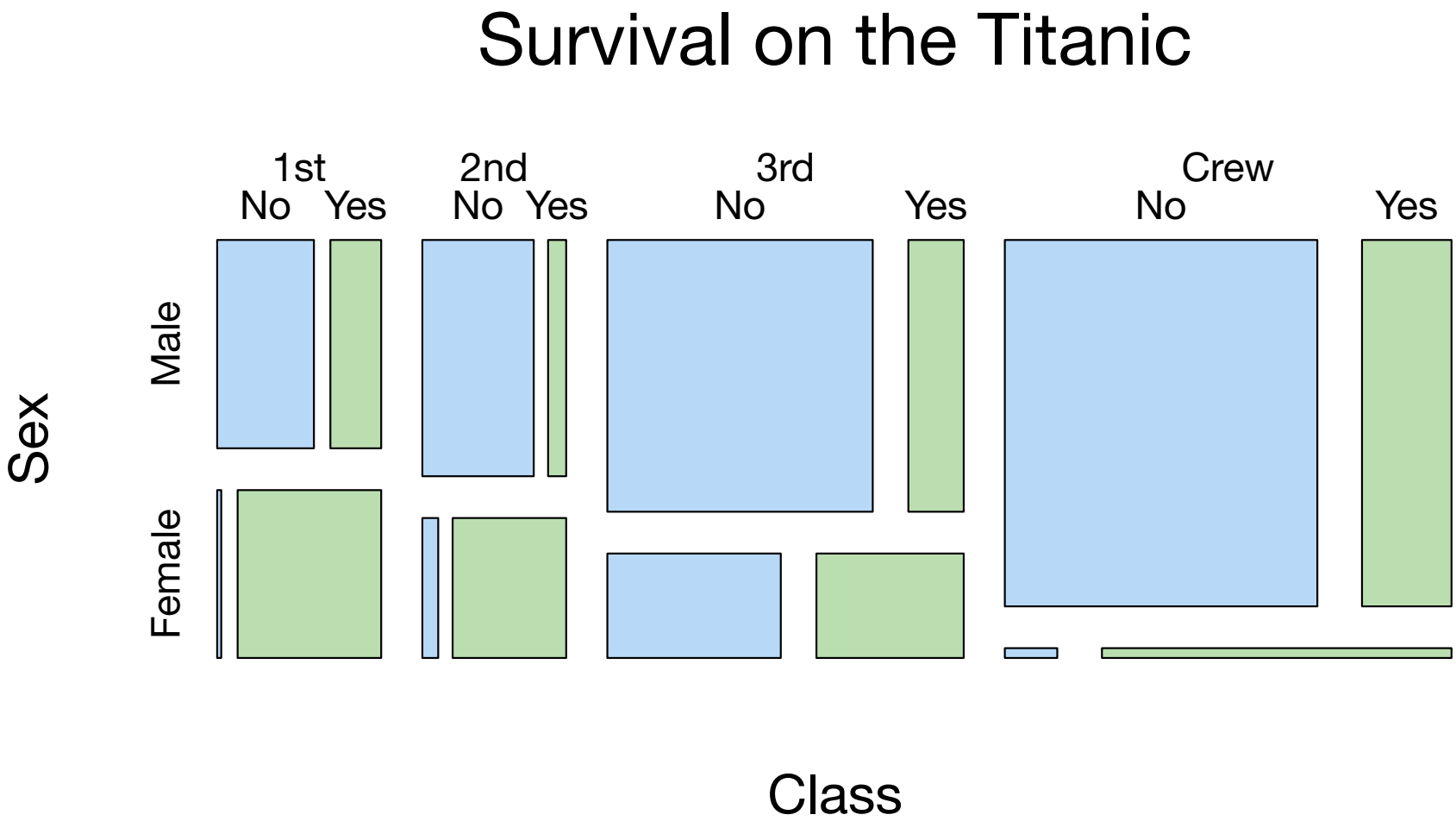


Abb.: Überleben (ja/nein), Geschlecht (männlich/weiblich) und Klasse (1/2/3/Besatzung) für $n = 2201$ Passagiere der Titanic

Zwei metrische Merkmale

- Es liegen Daten zu zwei intervall- oder verhältnisskalierten Merkmalen X und Y vor: Datenpaare $(x_i, y_i), \quad i = 1, \ldots, n$
  - Beispiel:
    X: Anzahl der Teilnahmen an der Übung in Statistik
    Y: Punkte in der Klausur in Statistik
- Frage: Gibt es einen Zusammenhang zwischen den Merkmalen X und Y ?

Streudiagramm

- Ein **Streudiagramm** ist eine Darstellung der Wertepaare $(x_1, y_1), \ldots, (x_n, y_n)$ im xy-Koordinatensystem.
- Man erhält eine Vorstellung über Streuung und Form der Punktwolke.
- Wird auch als *Scatterplot* bezeichnet.

Bsp. Streudiagramm (1)

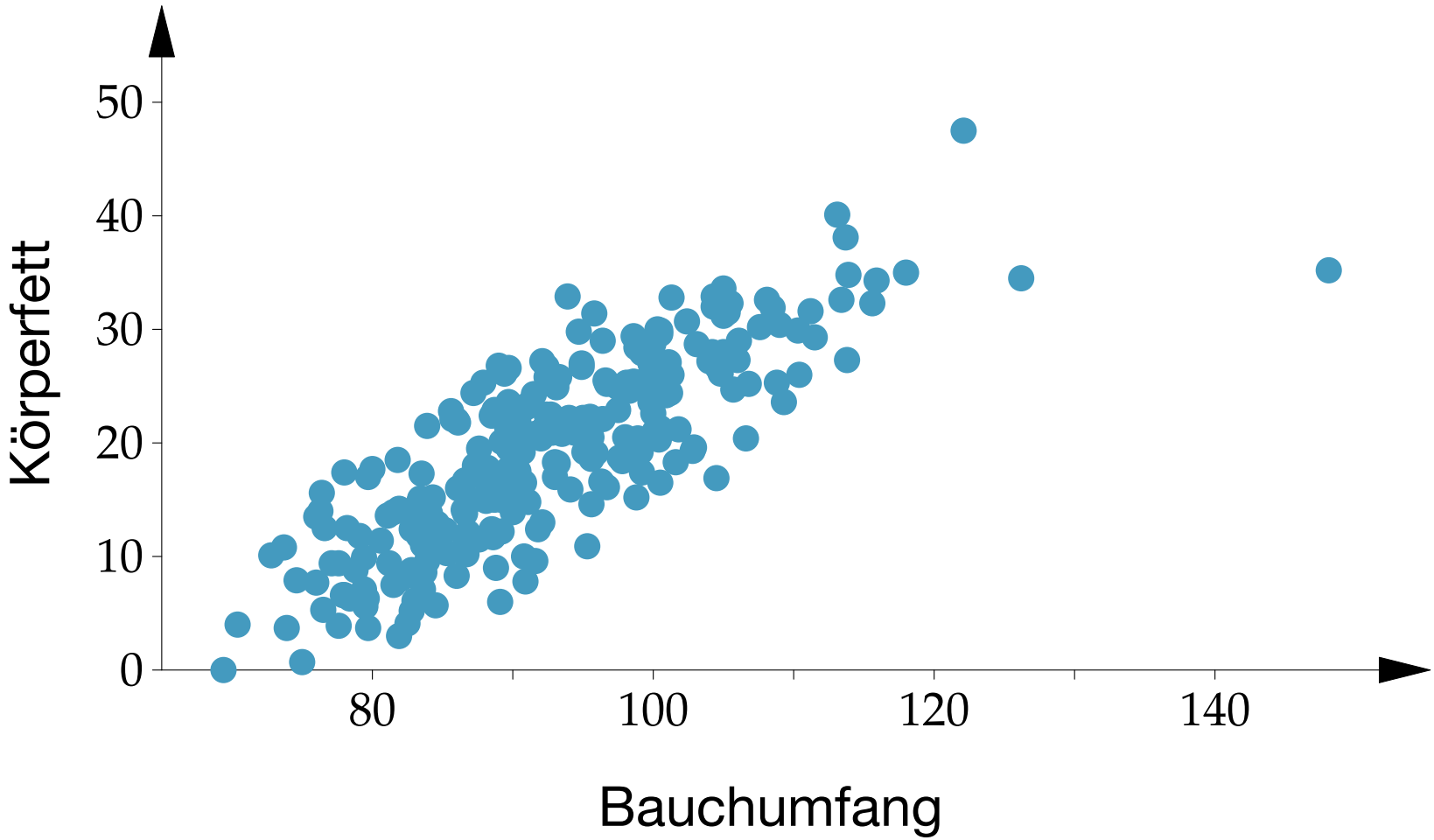


Abb.: Streudiagramm von Bauchumfang (in cm) und Körperfett (in Prozent) (Aus: Schlittgen 2012)

Abweichungsprodukte (1)

- Maß für den **wechselseitigen linearen Zusammenhang** zweier intervall- oder verhältnisskalierter Merkmale
- Idee: Beschreibung der Streuung einer beobachteten Punktwolke durch die Summe von Abweichungsprodukten

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

Bsp. Zwei intervall- oder verhältnisskalierte

i	x_i	y_i
1	3	1
2	1	2
3	8	9
4	7	7
5	2	3
6	4	9
7	10	11
$\sum$	35	42

Schwerpunkt der Punktwolke:

$(\bar{x}, \bar{y}) = (5, 6)$

Bsp. Streudiagramm (2)

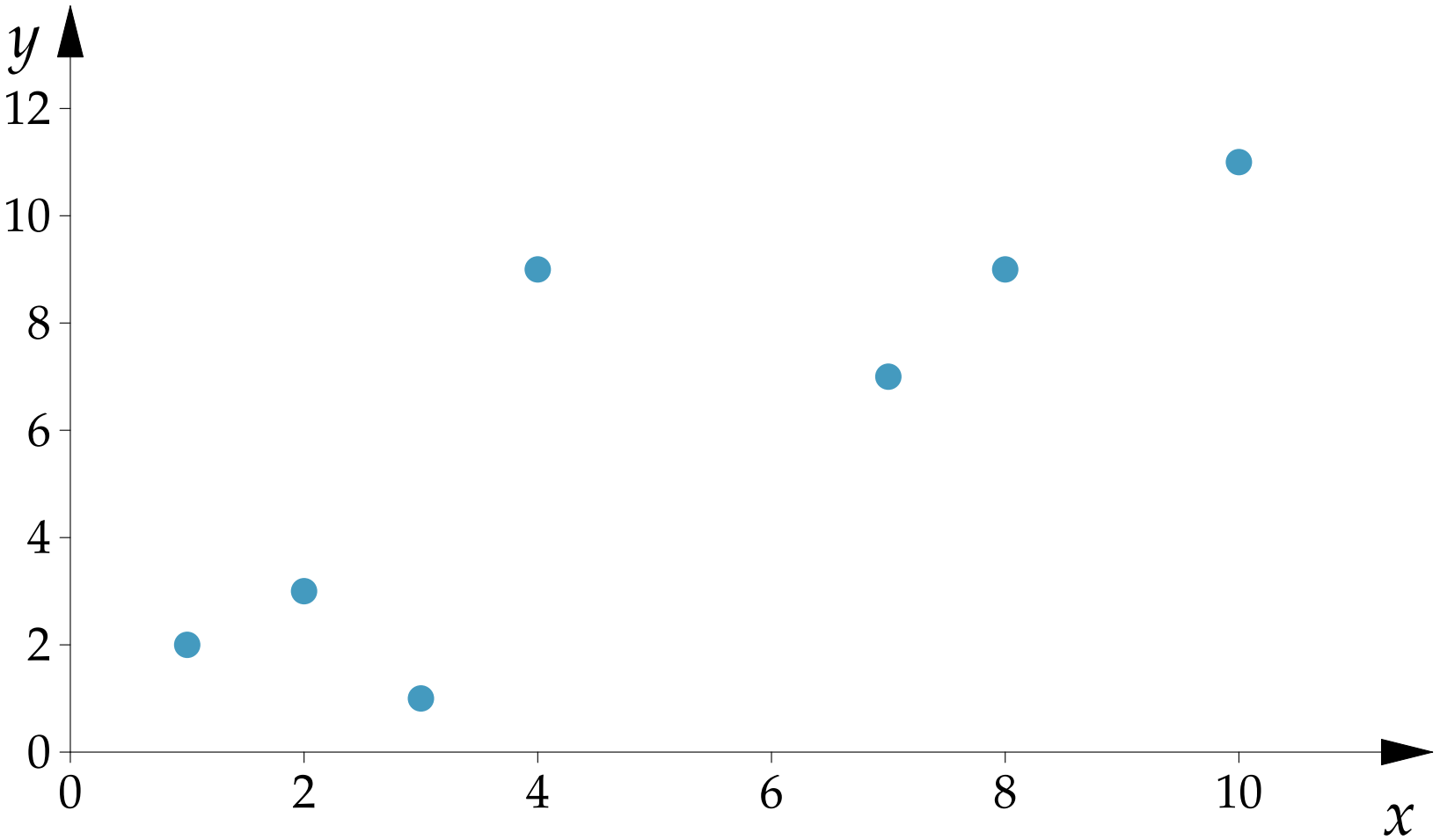


Abb.: Punktwolke der Merkmale X und Y

Bsp. Streudiagramm (3)

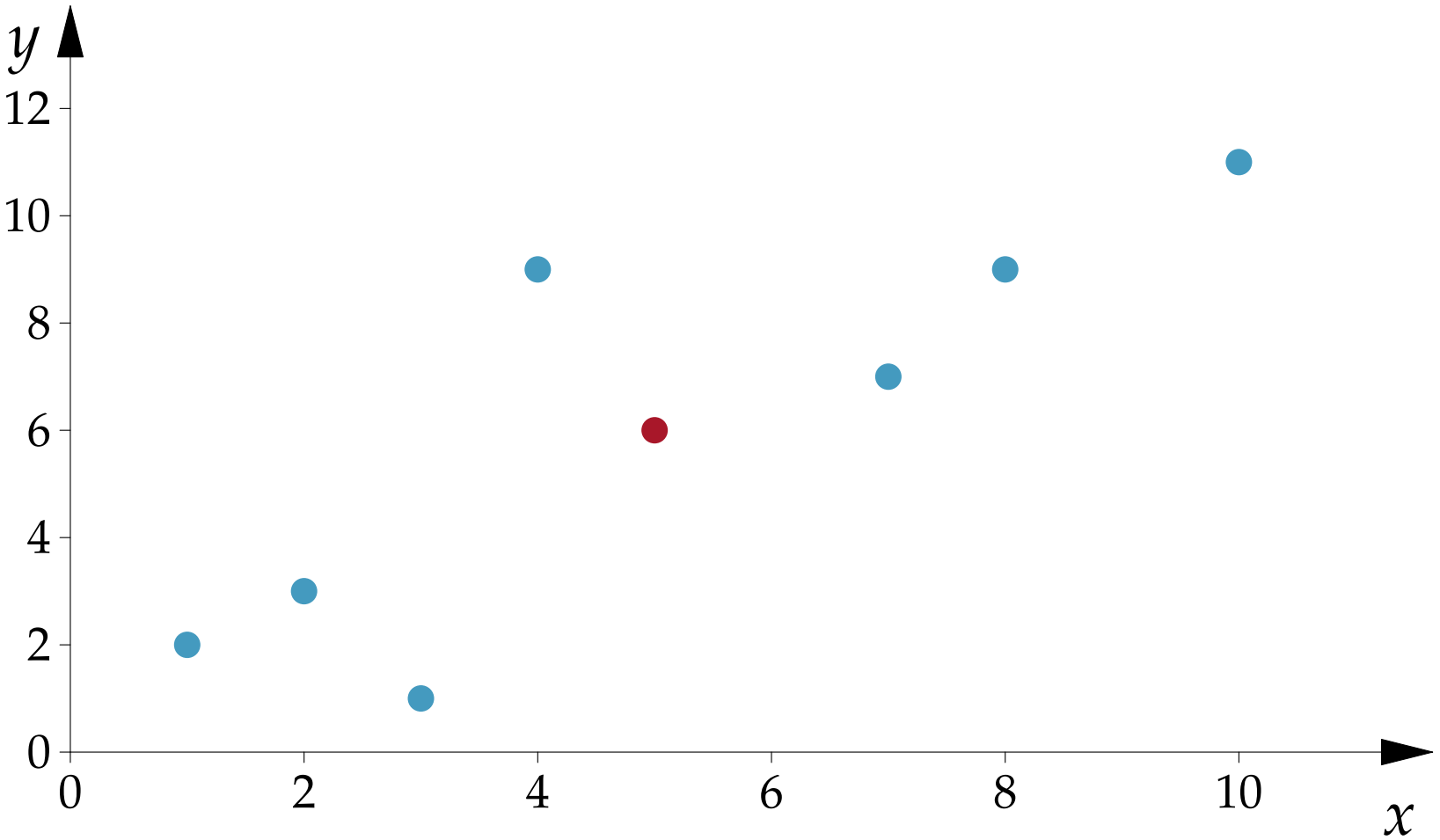


Abb.: Punktwolke (blau) zusammen mit dem Schwerpunkt (rot)

Bsp. Abweichungsprodukte (1)

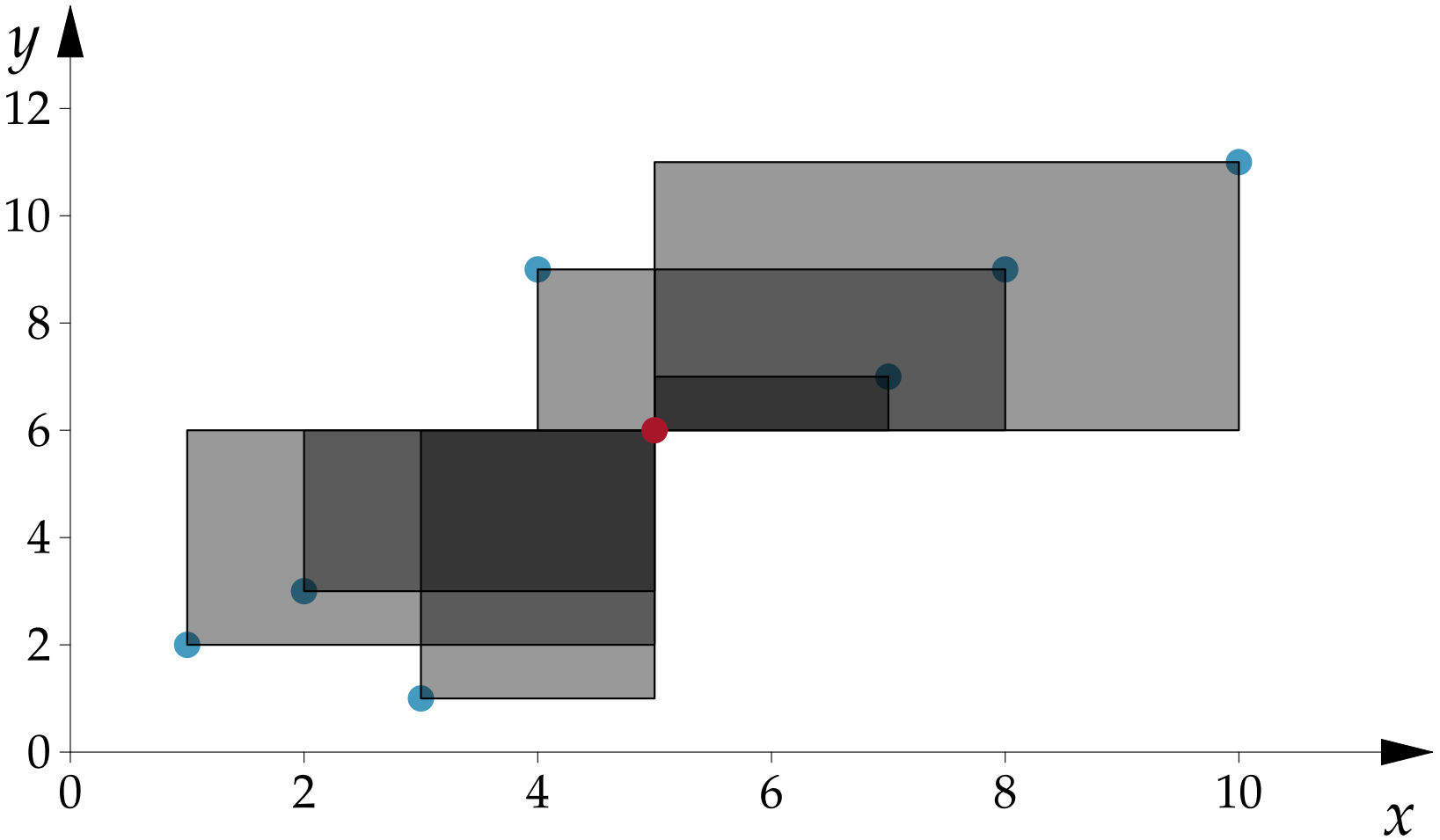


Abb.: Streudiagramm für Beispieldaten zusammen mit den Abweichungsprodukten (Rechtecke)

Abweichungsprodukte (2)

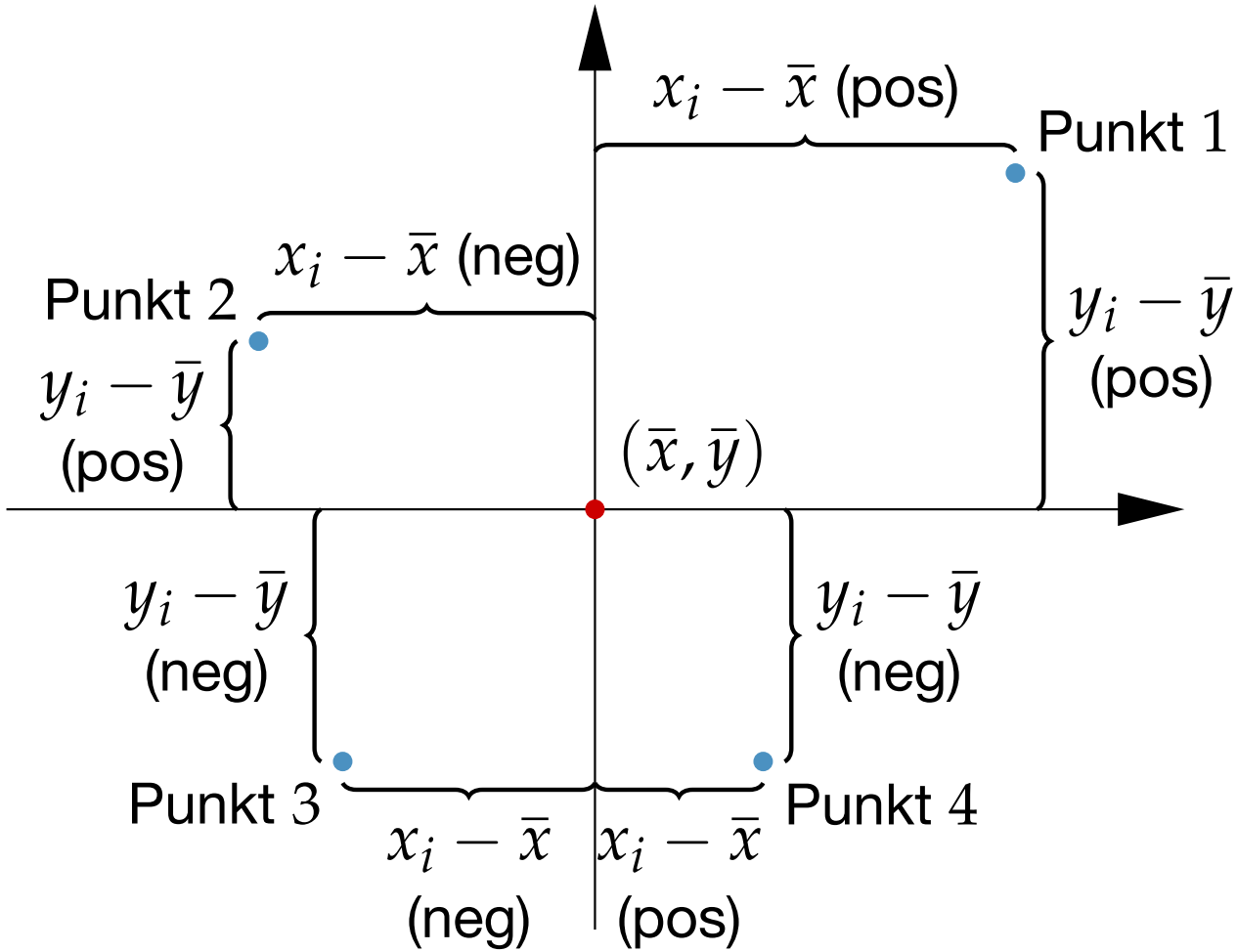


Abb.: Punkte im Koordinatensystem durch den Schwerpunkt $(\bar{x}, \bar{y})$

Abweichungsprodukte (3)

Tab.: Vorzeichen der einzelnen Komponenten und des Produkts der Abweichungen

	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
Punkt 1 (1. Quadrant)	positiv	positiv	positiv
Punkt 2 (2. Quadrant)	negativ	positiv	negativ
Punkt 3 (3. Quadrant)	negativ	negativ	positiv
Punkt 4 (4. Quadrant)	positiv	negativ	negativ

Bsp. Abweichungsprodukte (2)

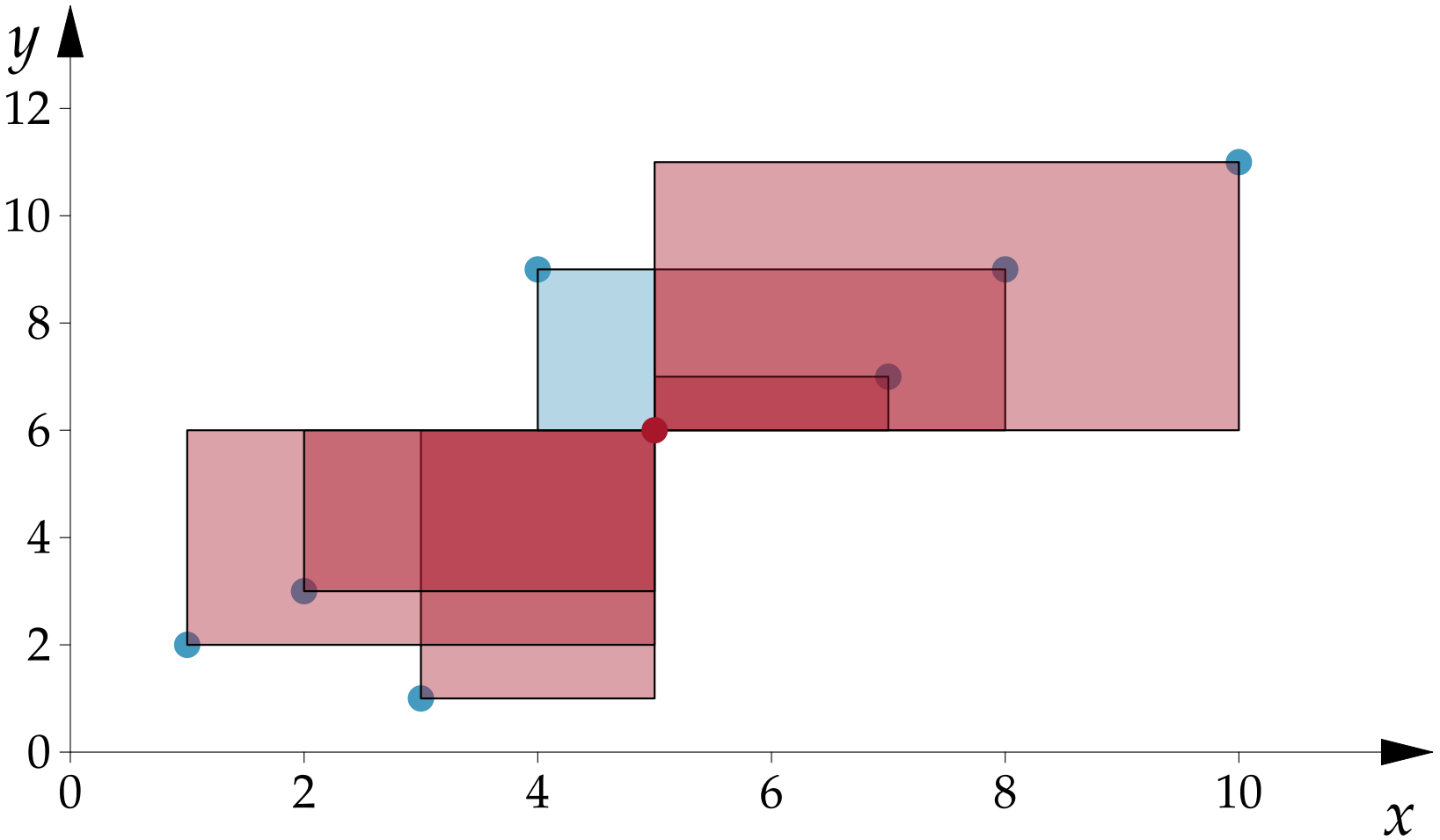


Abb.: Streudiagramm und Abweichungsprodukte (rot: positiv, blau: negativ)

Empirische Kovarianz (1)

- **empirische Kovarianz:** Summe der Abweichungsprodukte geteilt durch die Anzahl der Beobachtungspaare
  - Es gilt

$$\tilde{s}_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \,.$$

Bsp. Abweichungsprodukte (3)

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	3	1	-2	-5	10
2	1	2	-4	-4	16
3	8	9	3	3	9
4	7	7	2	1	2
5	2	3	-3	-3	9
6	4	9	-1	3	-3
7	10	11	5	5	25
$\sum$	35	42	0	0	68

Bsp. Abweichungsprodukte (4)

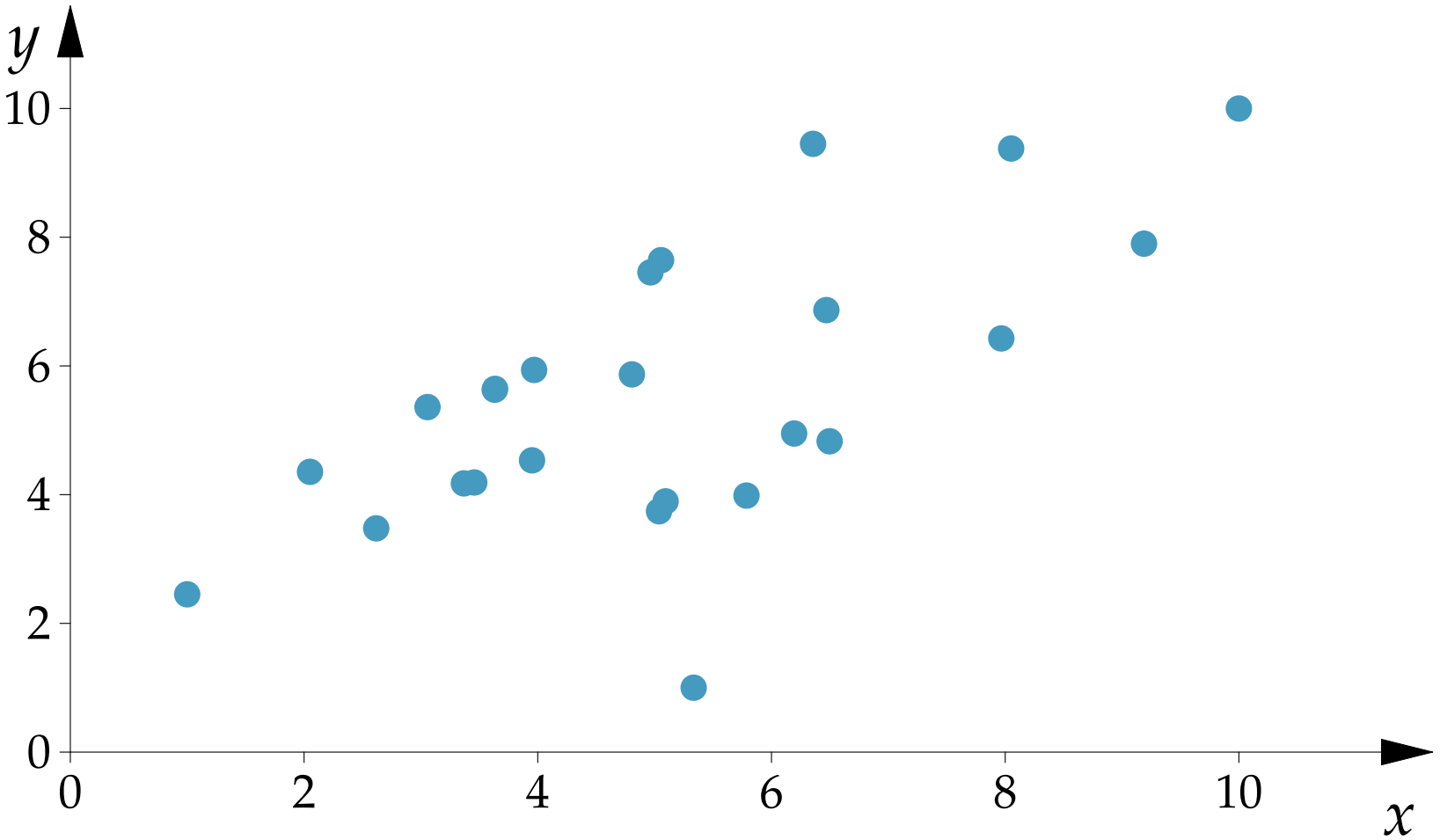


Abb.: Streudiagramm

Bsp. Abweichungsprodukte (5)

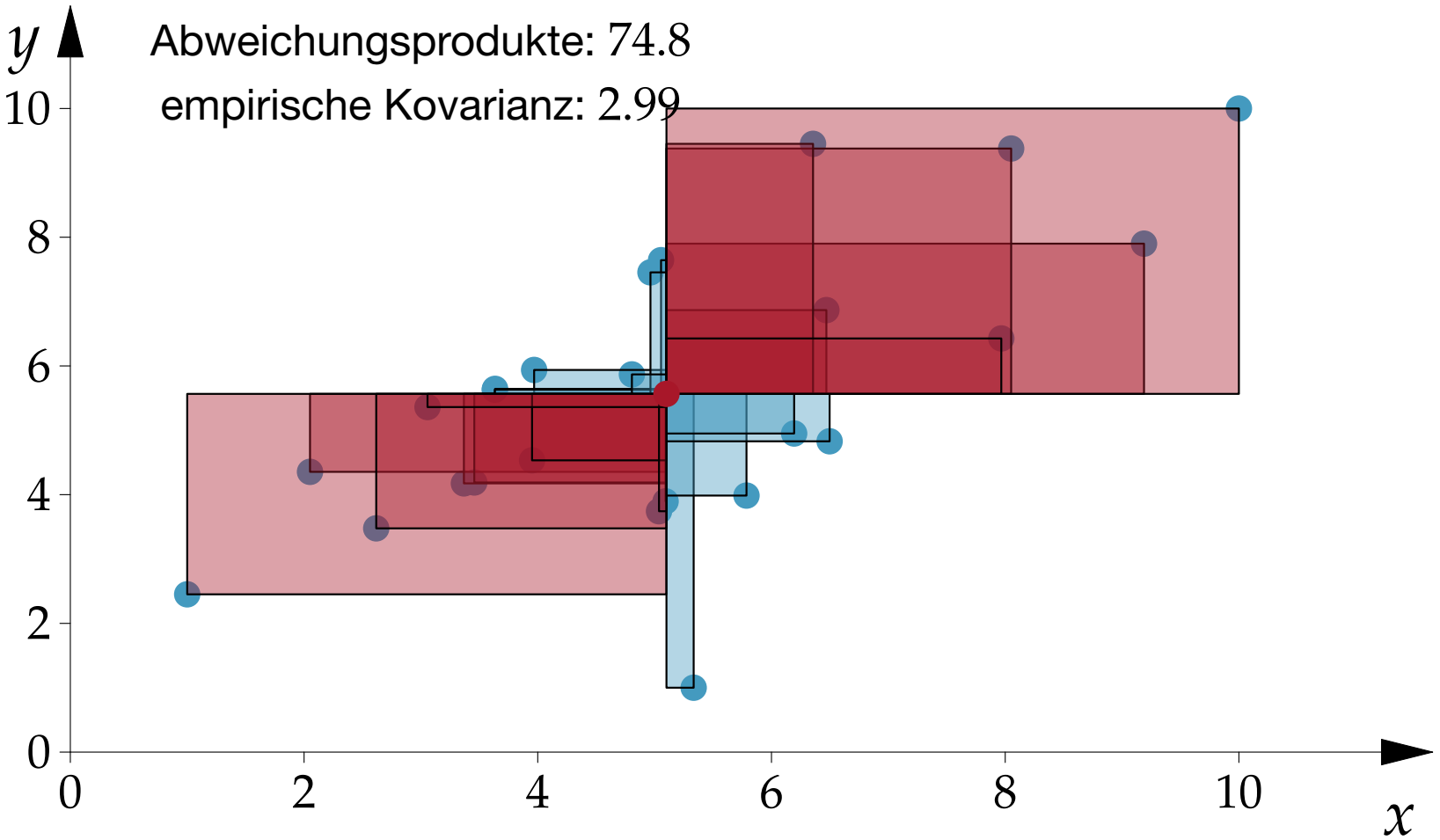


Abb.: Streudiagramm und Abweichungsprodukte (rot: positiv, blau: negativ)

Bsp. Abweichungsprodukte (6)

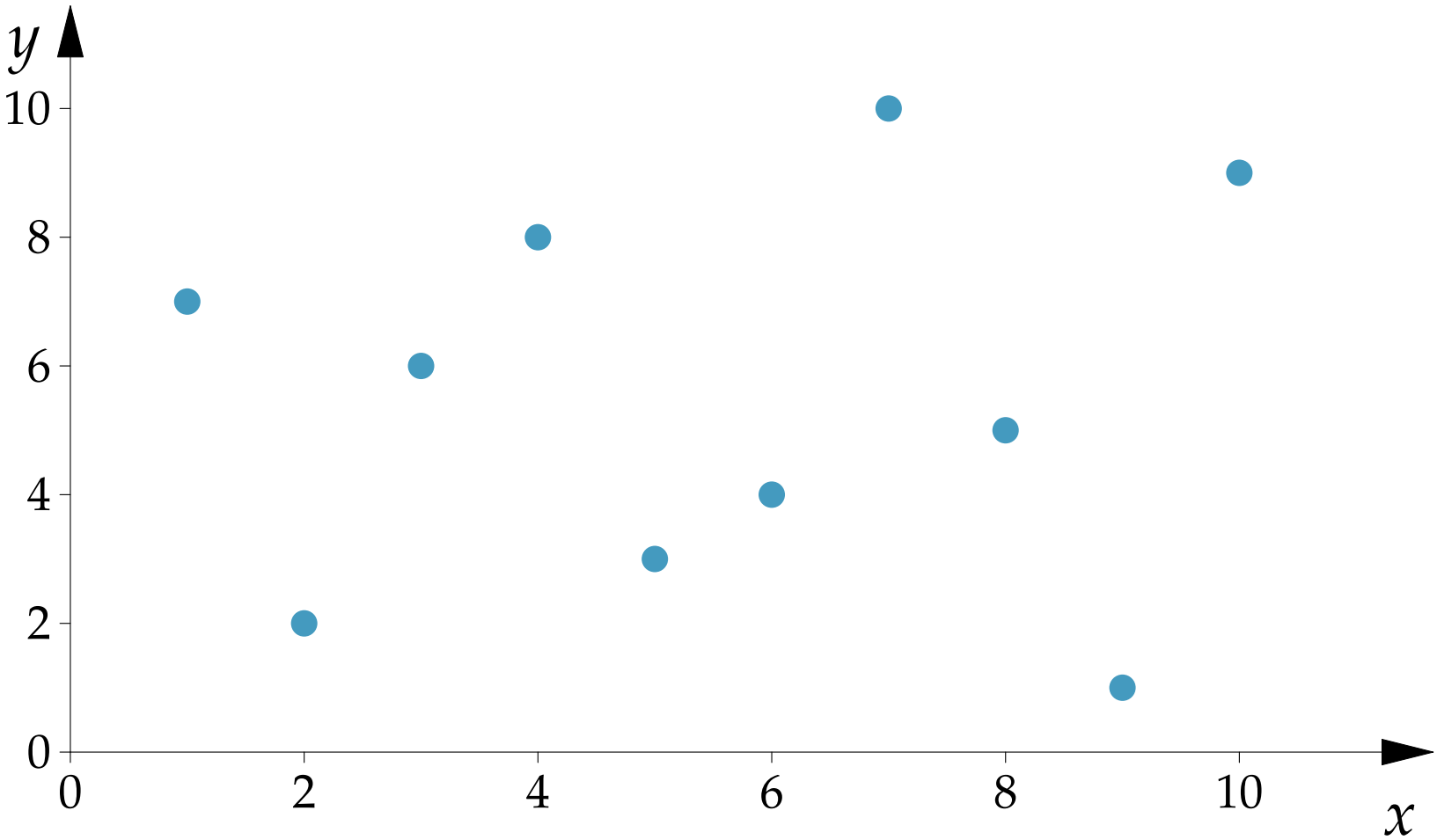


Abb.: Streudiagramm

Bsp. Abweichungsprodukte (7)

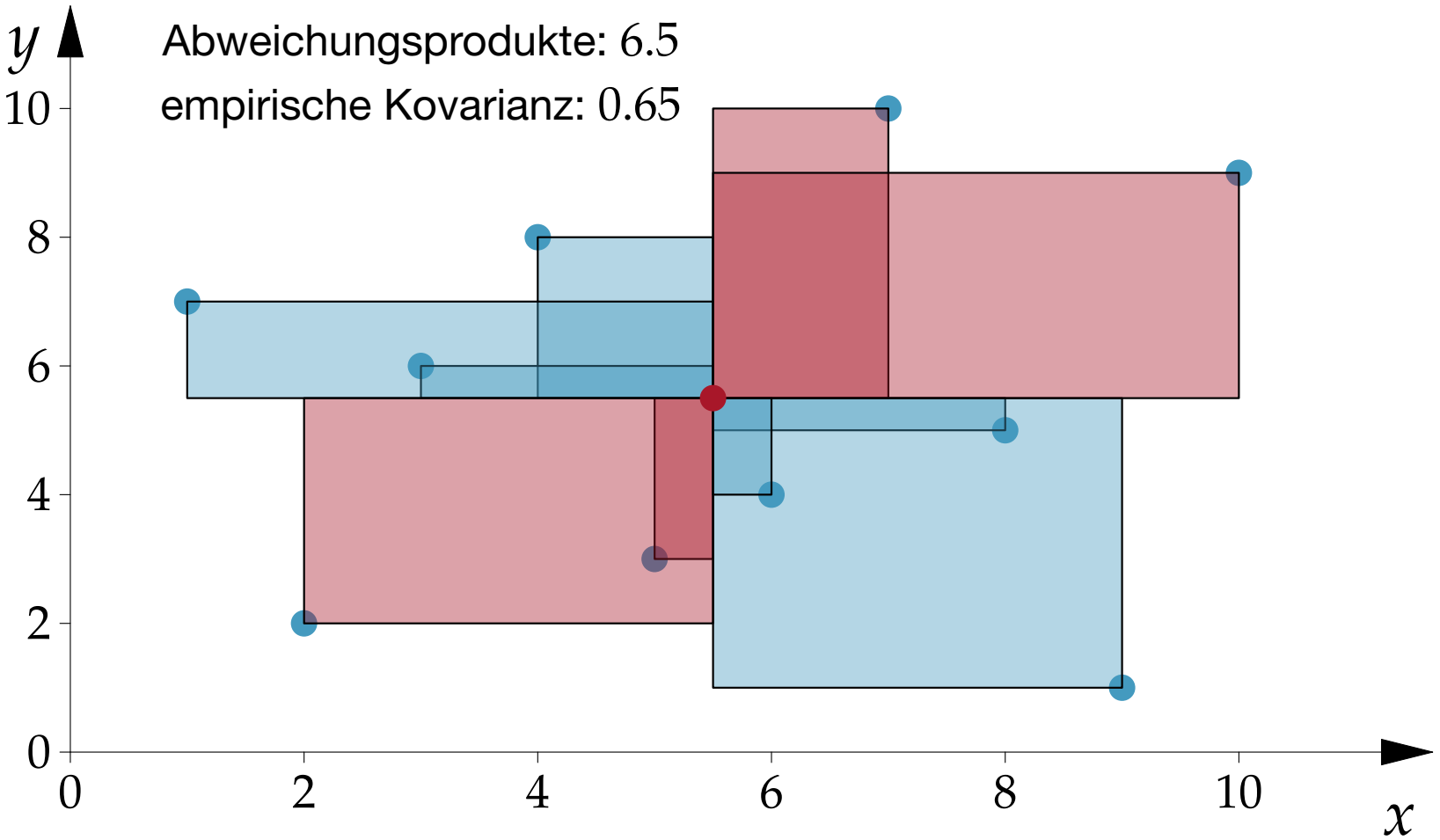


Abb.: Streudiagramm und Abweichungsprodukte (rot: positiv, blau: negativ)

Bsp. Abweichungsprodukte (8)

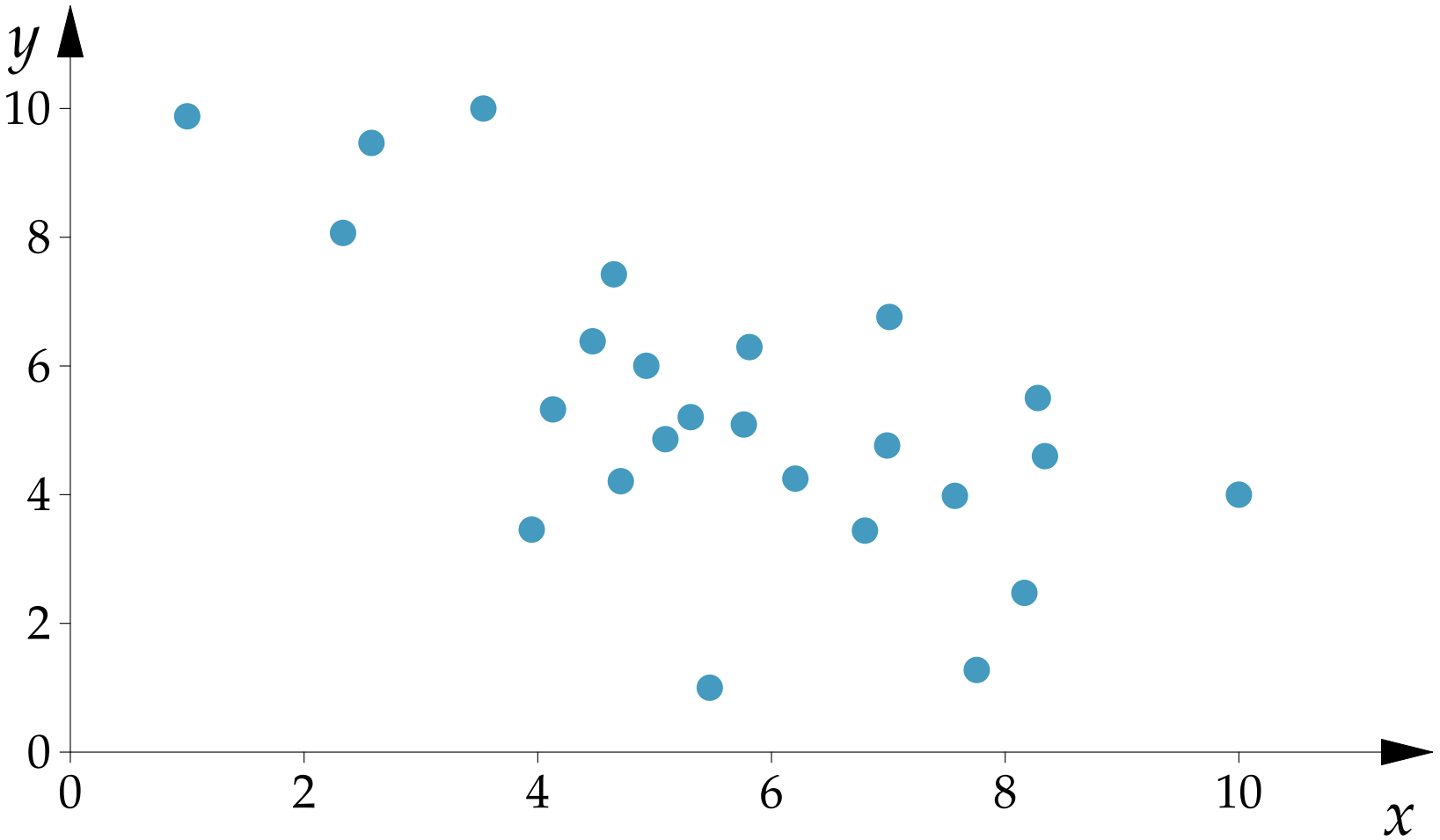


Abb.: Streudiagramm

Bsp. Abweichungsprodukte (9)

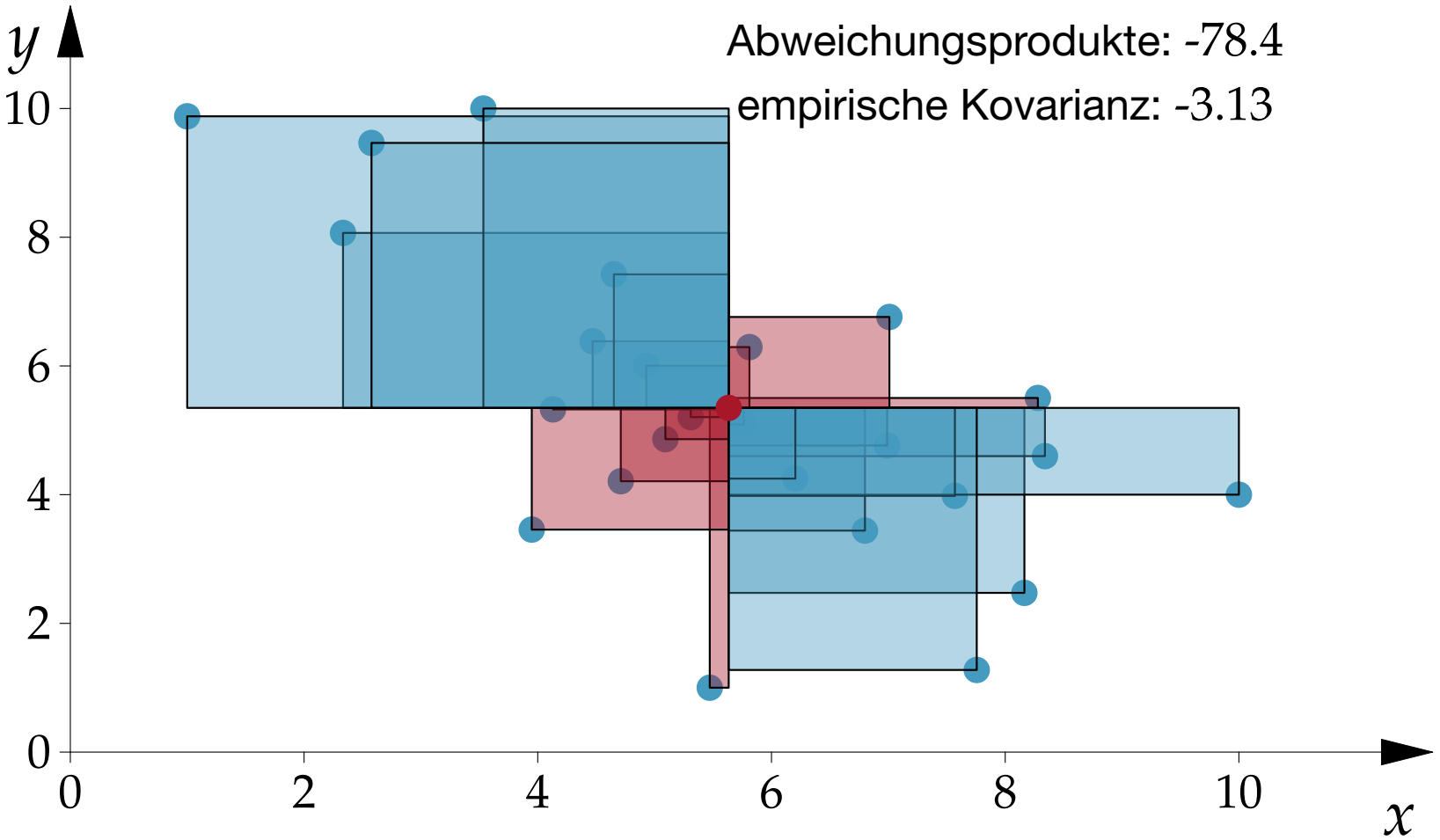


Abb.: Streudiagramm und Abweichungsprodukte (rot: positiv, blau: negativ)

Empirische Kovarianz (2)

- $\tilde{s}_{XY}$ ist positiv, wenn positive Abweichungsprodukte (Rechtecke) überwiegen.
- $\tilde{s}_{XY}$ ist negativ, wenn negative Abweichungsprodukte (Rechtecke) überwiegen.
- $\tilde{s}_{XY}$ ist 0, wenn beide Anteile gleich groß sind.
- Für $\tilde{s}_{XX}$ ergibt sich die empirische Varianz von X.
- Die Kovarianz s_{XY} ergibt sich für die Mittlung mit $\frac{1}{n-1}$ anstatt $\frac{1}{n}$.
- Die empirische Kovarianz sowie die Kovarianz sind nicht invariant gegenüber Maßstabsveränderungen, d. h. sie sind nicht normiert.

Bsp. Maßstabsveränderungen

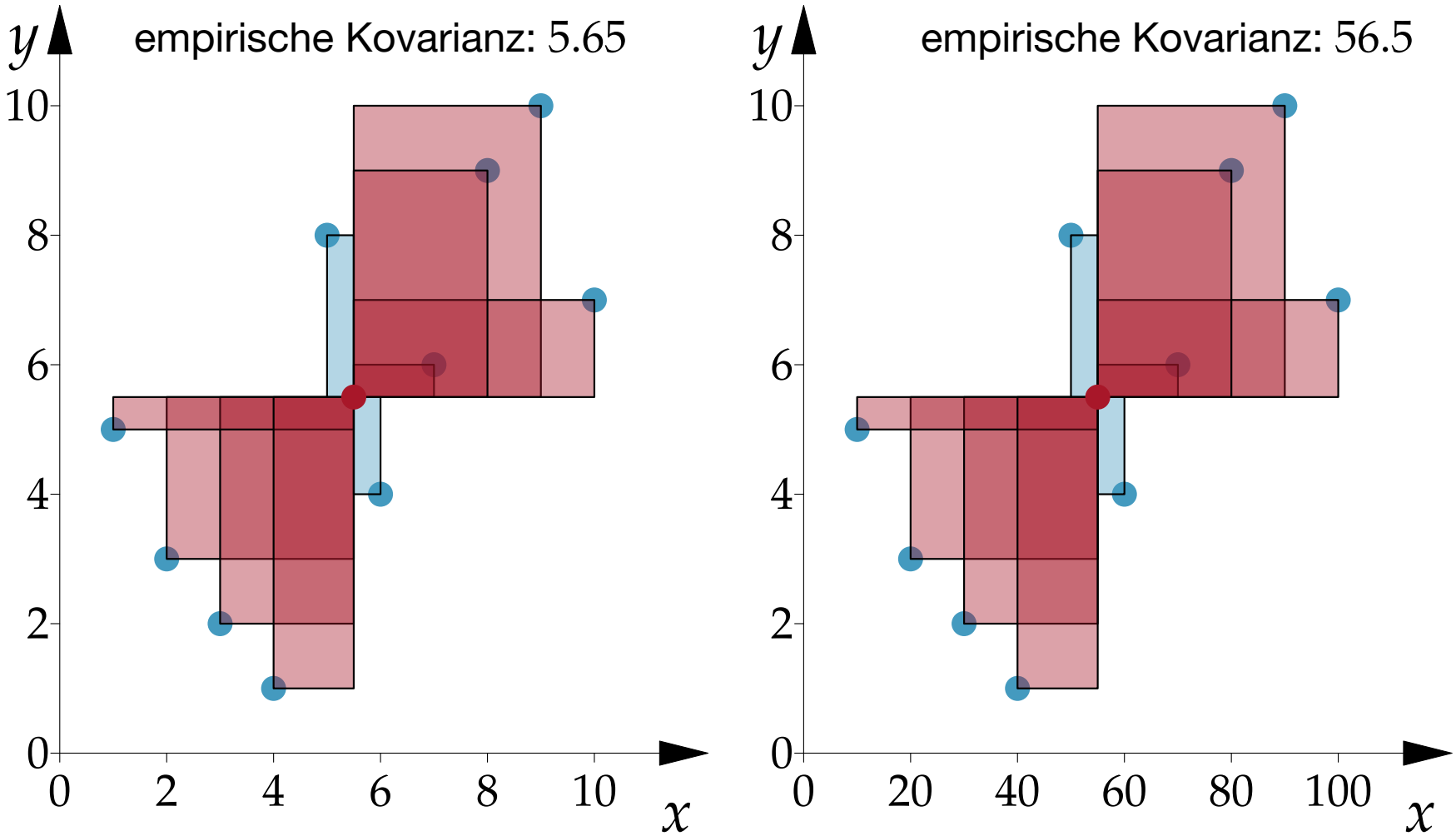


Abb.: Streudiagramme und Abweichungsprodukte für unterschiedlich skalierte x-Achsen

Emp. Korrelationskoeffizient (1)

- **empirischer Korrelationskoeffizient:** die durch das Produkt der Standardabweichungen normierte empirische Kovarianz
  - Für die Daten (x_i, y_i), $i = 1, \ldots, n$ gilt

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\tilde{s}_{XY}}{\sqrt{\tilde{s}_{XX}} \sqrt{\tilde{s}_{YY}}}.$$

  - Der Wertebereich ist $-1 \leq r \leq 1$.
  - Dies ist nur für intervall- und verhältnisskalierte Merkmale sinnvoll definiert.
- auch **Bravais-Pearson-Korrelationskoeffizient** genannt

Wdh. Bsp. Streudiagramm (3)

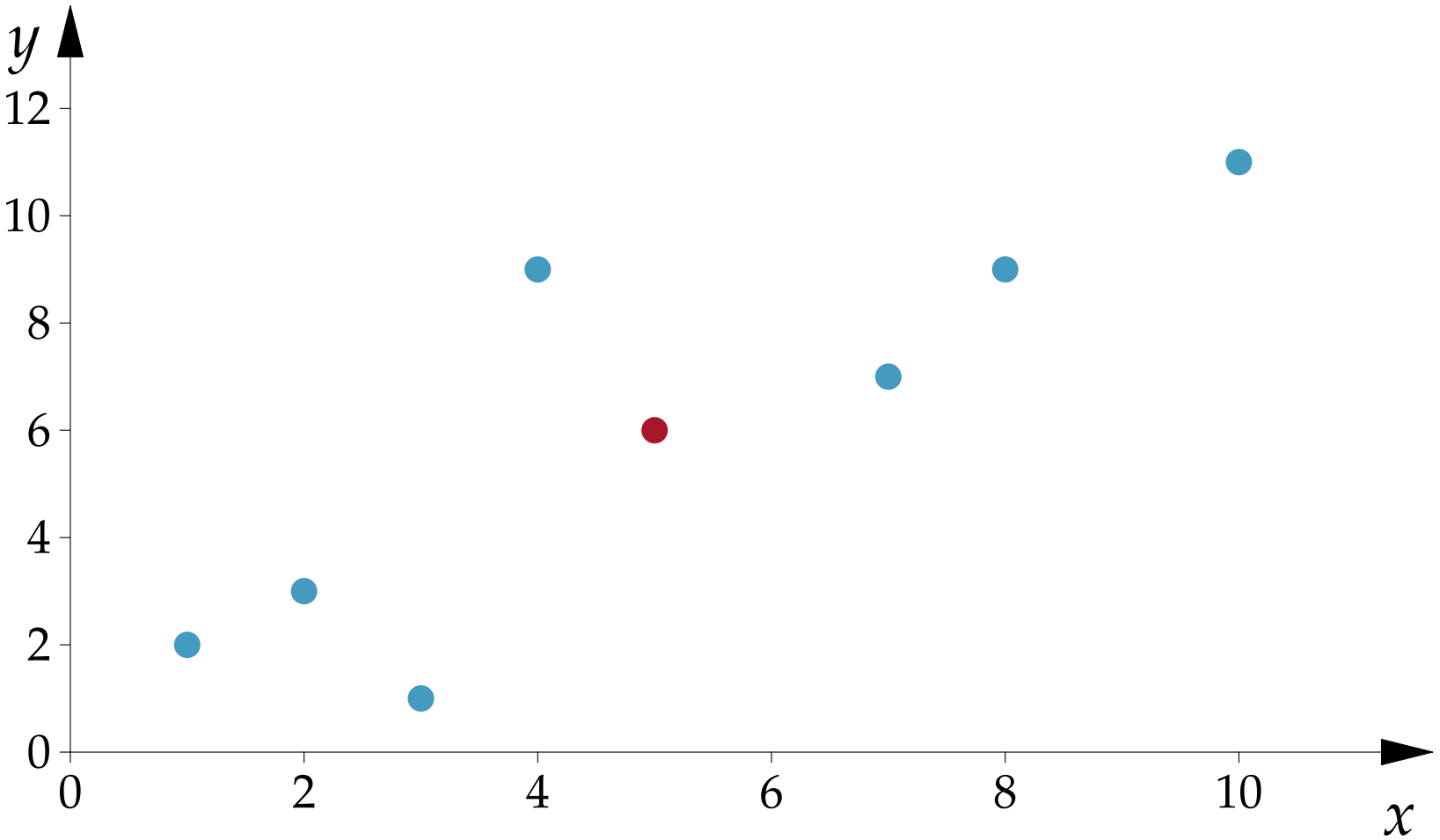


Abb.: Punktwolke (blau) zusammen mit dem Schwerpunkt (rot)

Bsp. Korrelationskoeffizient

i	x_i	y_i	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	3	1	10	4	25
2	1	2	16	16	16
3	8	9	9	9	9
4	7	7	2	4	1
5	2	3	9	9	9
6	4	9	-3	1	9
7	10	11	25	25	25
Σ	35	42	68	68	94
$\bar{x}$	5				
$\bar{y}$		6			

Emp. Korrelationskoeffizient (2)

- $r > 0$: positiver linearer Zusammenhang,
  Tendenz: Werte (x_i, y_i) um eine Gerade positiver Steigung liegend
- $r < 0$: negativer linearer Zusammenhang,
  Tendenz: Werte (x_i, y_i) um eine Gerade negativer Steigung liegend
- $r = 0$: kein linearer Zusammenhang
- $r = 1$, $(r = -1)$: alle Punkte auf einer Geraden mit positiver (negativer) Steigung liegend

Bsp. Emp. Korrelationskoeffizient (1)

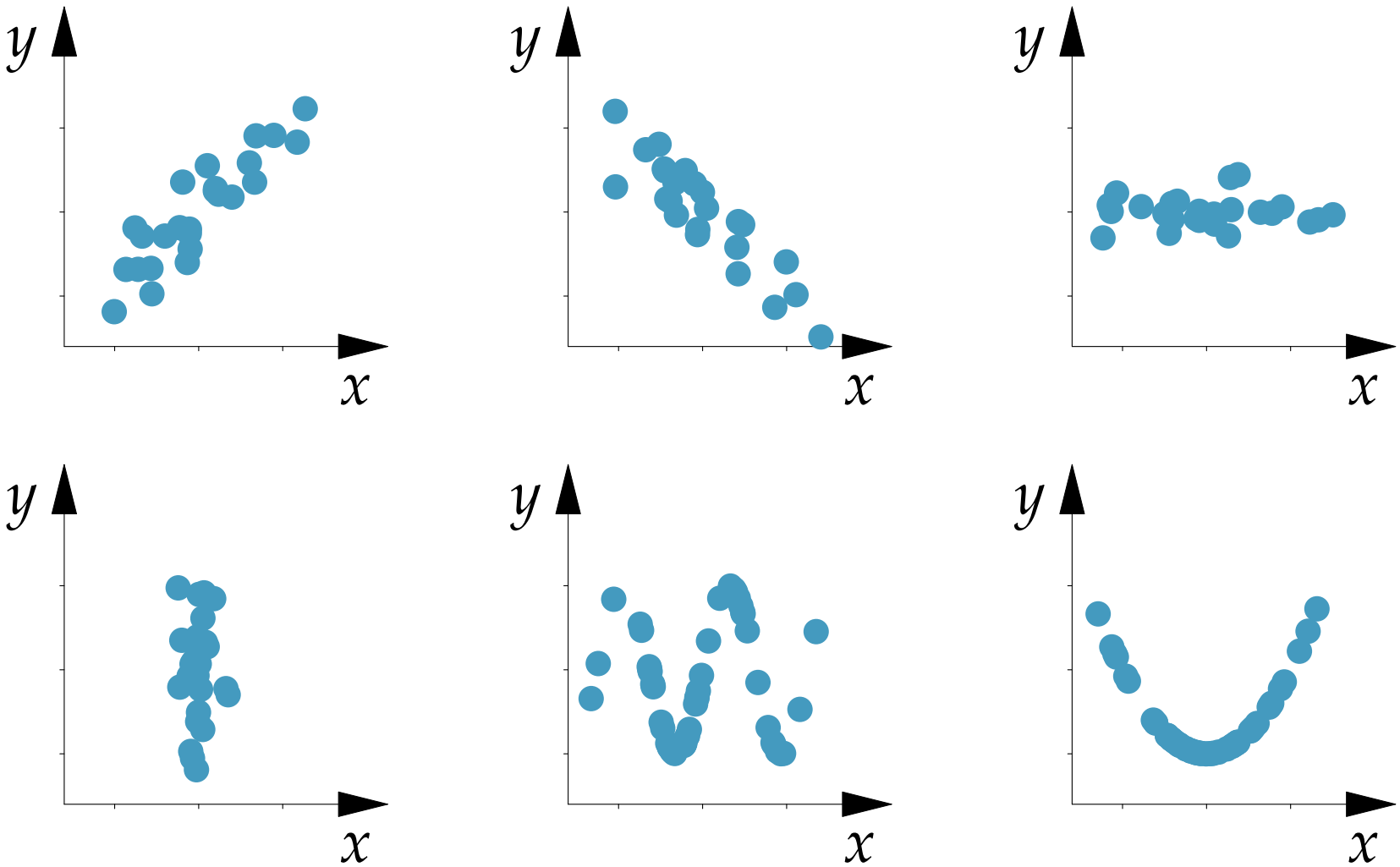


Abb.: Punktwolken und Korrelationskoeffizienten

Bsp. Emp. Korrelationskoeffizient (2)

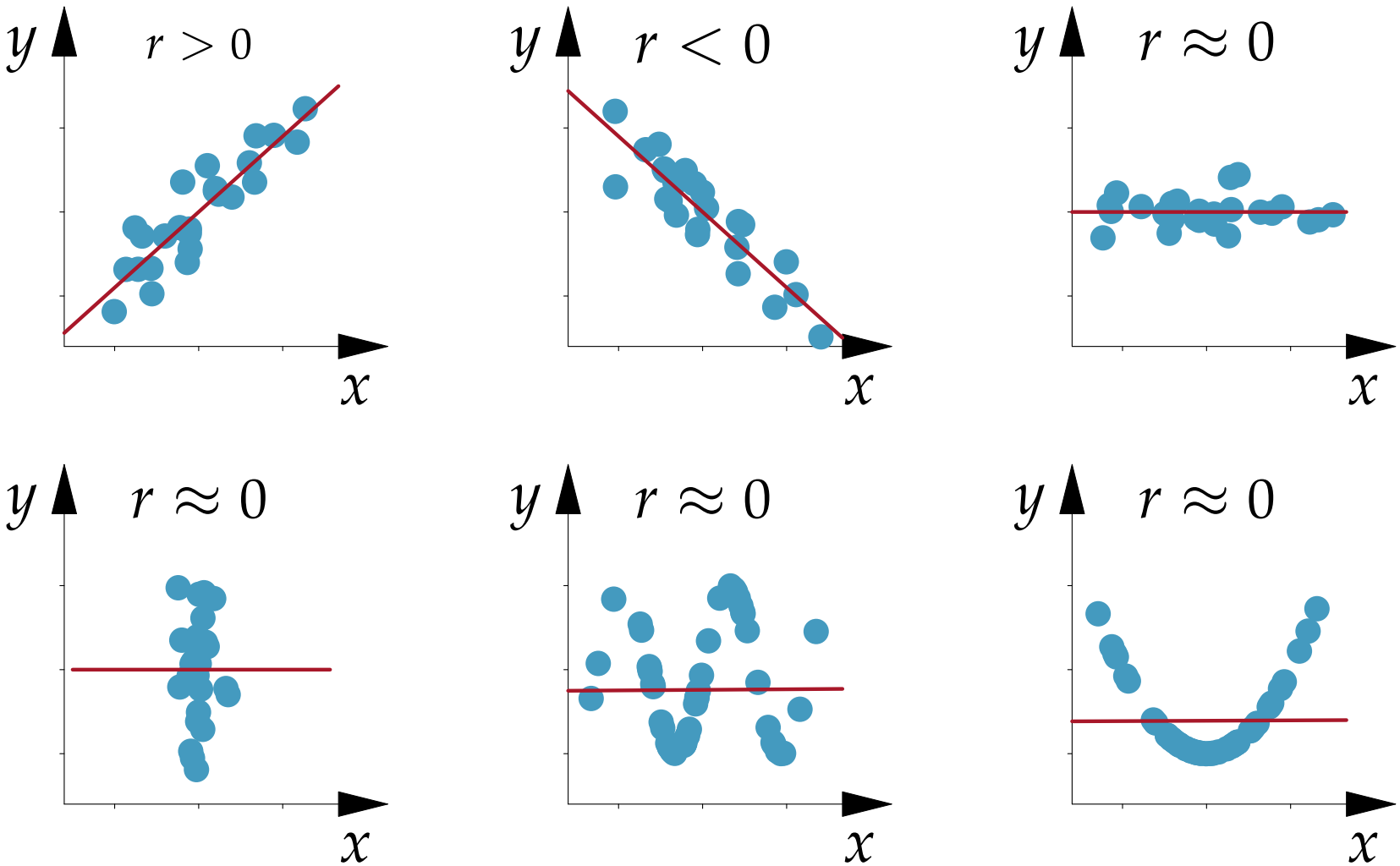


Abb.: Punktwolken und Korrelationskoeffizienten

Bsp. Emp. Korrelationskoeffizient (3)

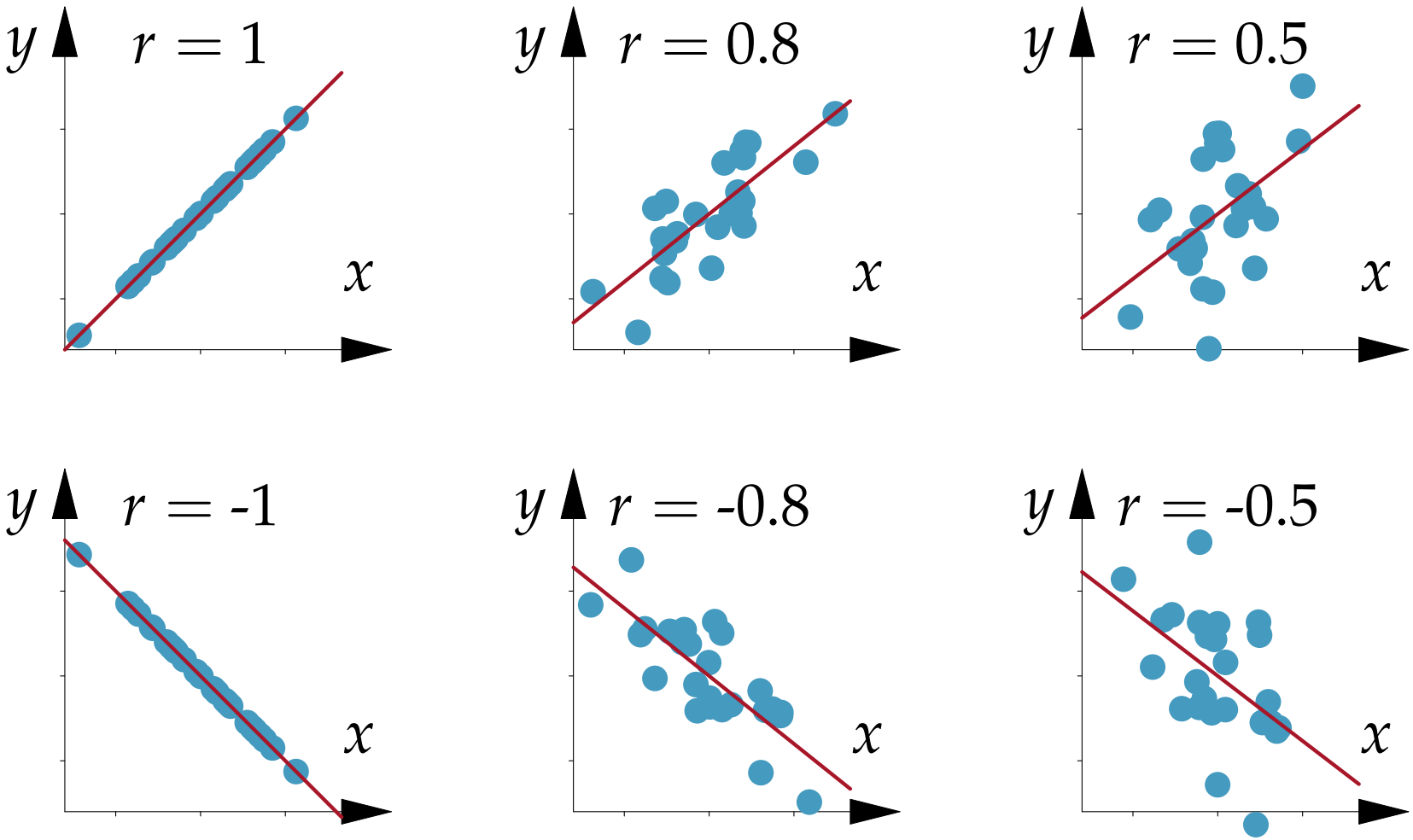


Abb.: Punktwolken und Korrelationskoeffizienten

Bsp. Emp. Korrelationskoeffizient (4)

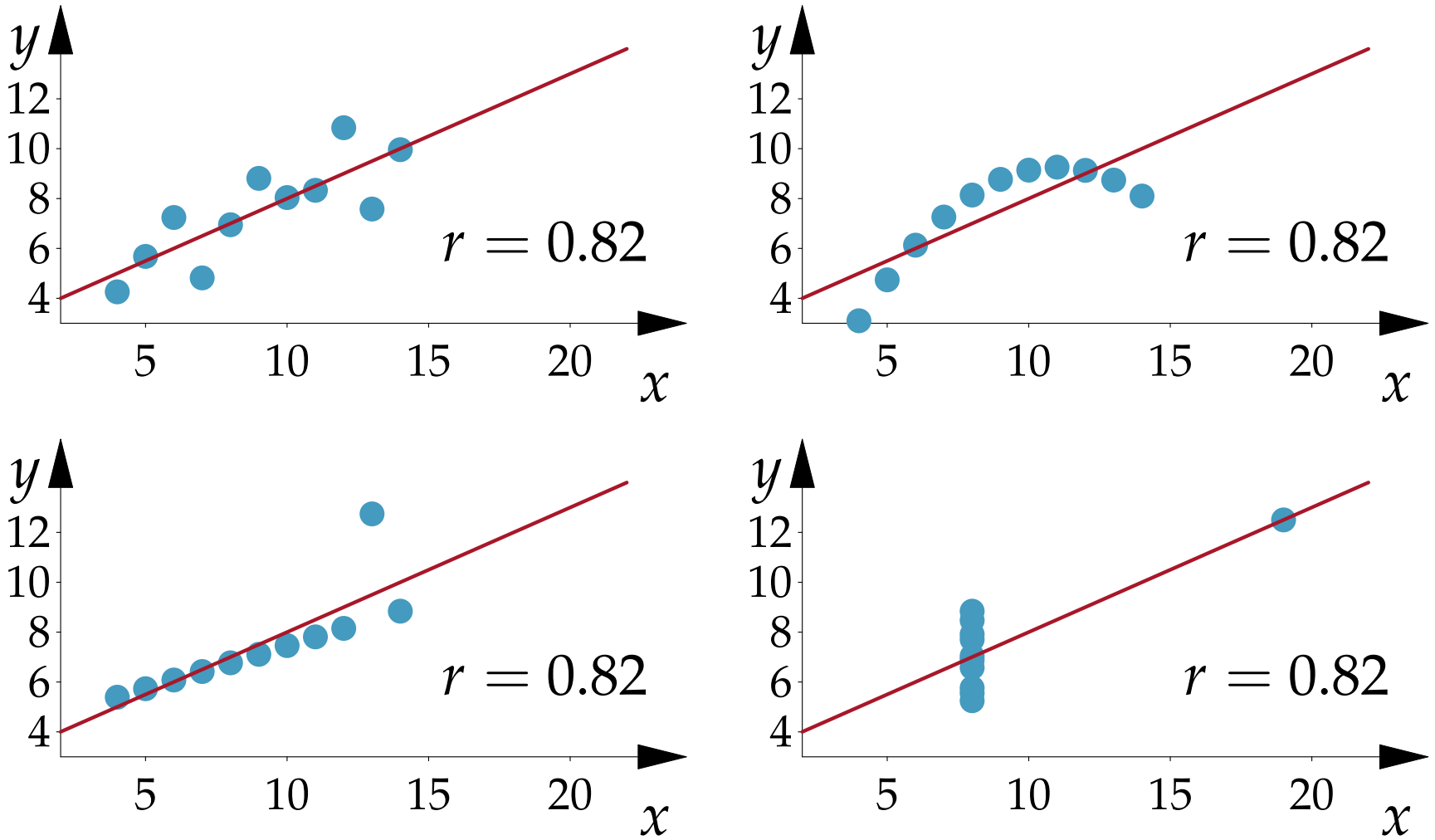
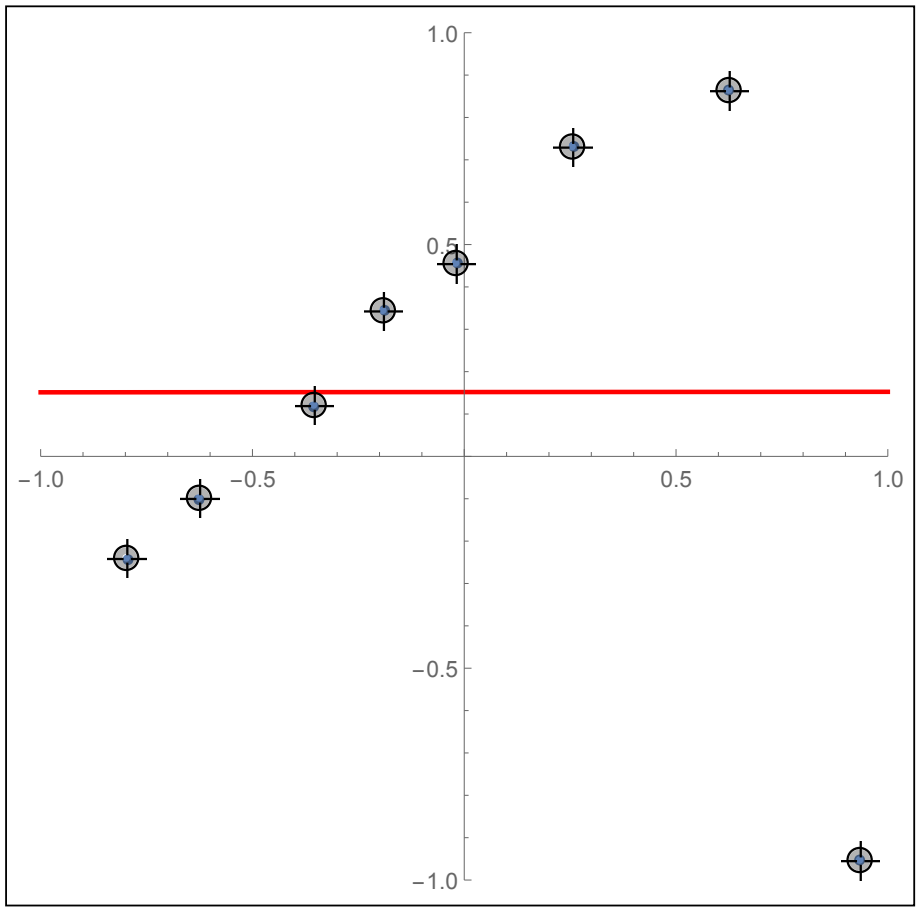


Abb.: Unterschiedliche Punktwolken mit gleichem Korrelationskoeffizienten (Aus: Anscombe, 1973)

Bem. Emp. Korrelationskoeffizient

- Faustregel zur Stärke des linearen Zusammenhangs ist:
  - schwache Korrelation: $|r| < 0.5$
  - mittlere Korrelation: $0.5 \leq |r| < 0.8$
  - starke Korrelation: $0.8 \leq |r|$
- Korrelationskoeffizienten sind insbesondere hilfreich als *vergleichende* Maße.
- Der Korrelationskoeffizient nach Pearson reagiert sensibel auf Ausreißer.

Bsp. Ausreißer



Pearsons r: 0.001

Abb.: Beispiel für die Empfindlichkeit des Korrelationskoeffizienten nach Pearson gegenüber Ausreißern