Rangkorrelationskoeffizient

- Idee: Korreliere nicht die Datenwerte selbst miteinander, sondern ihre unabhängig voneinander ermittelten Ränge.

Rangdaten

- Der **Rang eines Datenwertes** ist die Zahl, die angibt an welcher Stelle in der größensortierten Liste sich der Datenpunkt befindet.
  - Bezeichnen $x_{(1)} \leq \ldots \leq x_{(n)}$ die geordneten x-Werte, dann gilt

    $$rg(x_{(i)}) = i$$

    bzw. mit den geordneten y-Werten

    $$rg(y_{(i)}) = i\,.$$

- Damit ergeben sich aus den ursprünglichen Messpaaren $(x_i, y_i), i = 1, \ldots, n$ die neuen Rangdaten $(rg(x_i), rg(y_i)), i = 1, \ldots, n$.

Bindungen

- Falls alle Beobachtungswerte unterschiedlich sind, kommt jeder Rangwert genau einmal vor.
- Treten dem Wert nach gleiche Beobachtungen auf, so spricht man von **Bindungen**.
- Gleichen Beobachtungen wird der arithmetische Mittelwert der Ränge zugewiesen.

Bsp. Streudiagramm für Daten

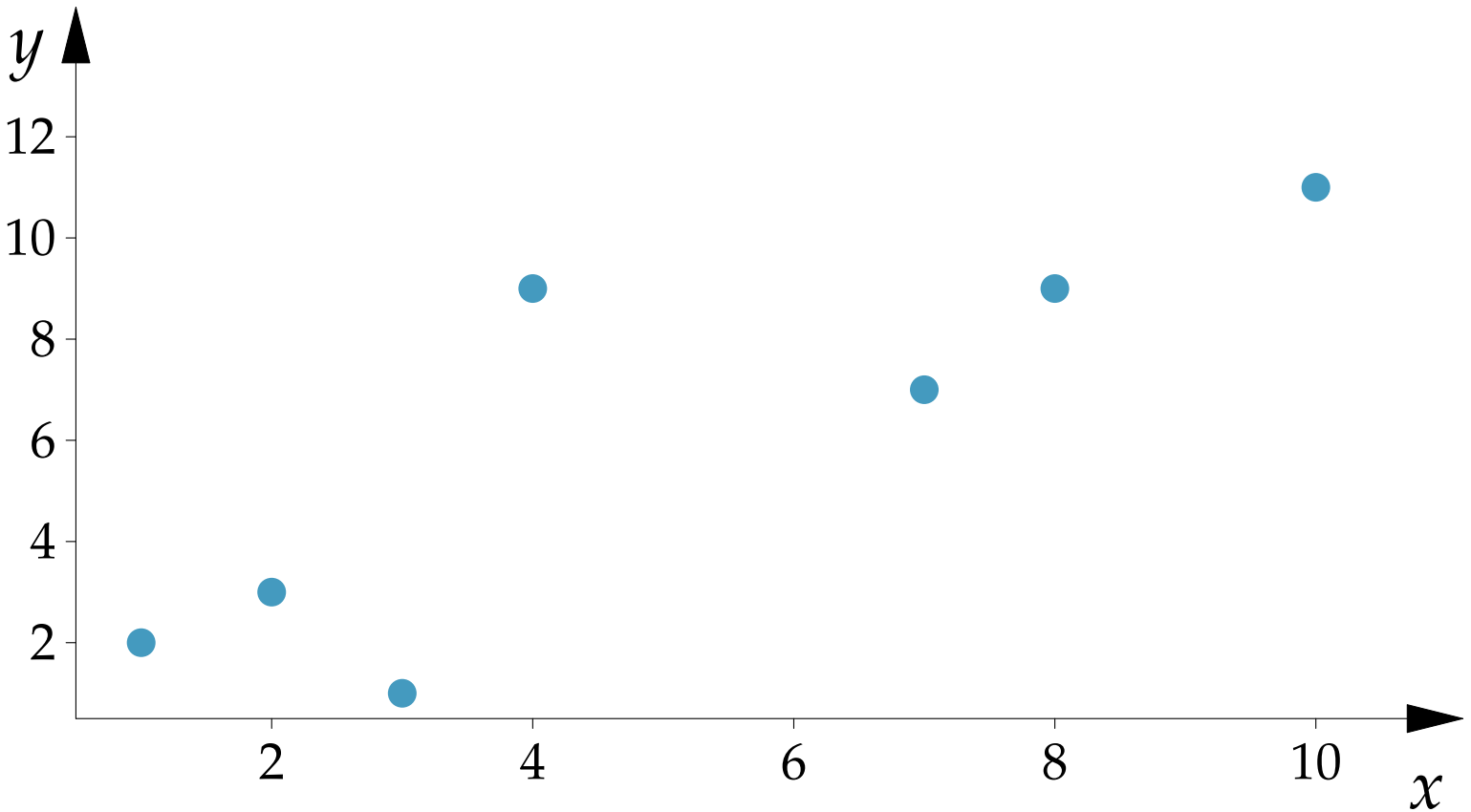


Abb.: Punktwolke der Daten

Bsp. Streudiagramm für Rangdaten

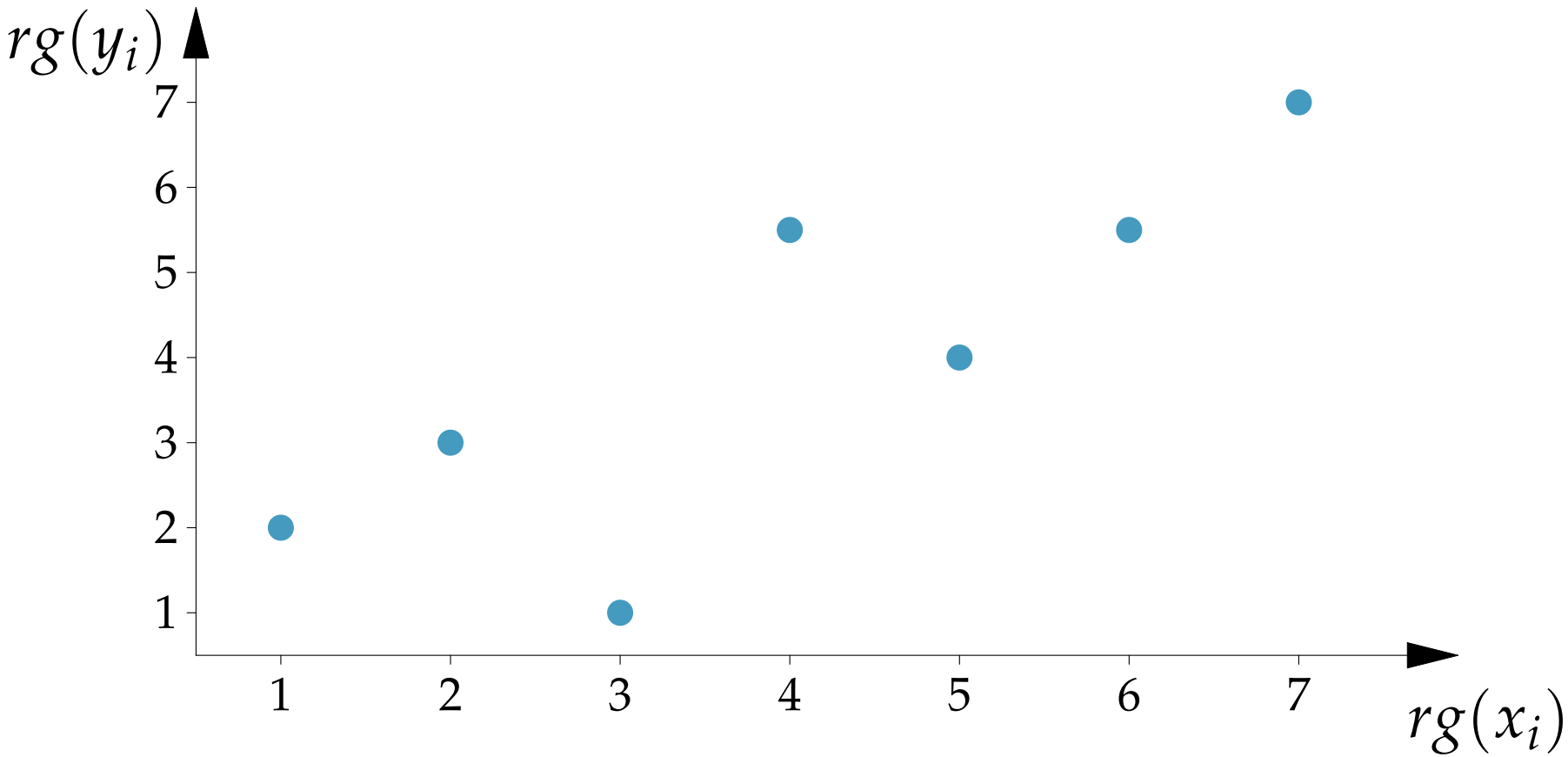


Abb.: Punktwolke der Rangdaten

Rangkorrelationskoeffizient (1)

- **Rangkorrelationskoeffizient:**
  Pearson-Korrelationskoeffizient, angewandt auf Rangdaten
  - Es gilt für die Rangpaare $(rg(x_i), rg(y_i))$, $i = 1, \ldots, n$

$$r_{\mathrm{SP}} = \frac{\sum (rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)}{\sqrt{\sum (rg(x_i) - \overline{rg}_X)^2} \sqrt{\sum (rg(y_i) - \overline{rg}_Y)^2}} \,.$$

  - Der Wertebereich ist $-1 \leq r_{\mathrm{SP}} \leq 1$.
  - Ist nur für mindestens ordinalskalierte Merkmale anwendbar.
- auch als **Spearmans Korrelationskoeffizient** bezeichnet

Rangkorrelationskoeffizient (2)

- Die Rangsummen sind gegeben durch

$$\sum_{i=1}^{n} rg(x_i) = \sum_{i=1}^{n} rg(y_i) = 1 + 2 + \ldots + n$$

$$= \frac{n(n+1)}{2}.$$

Division durch n ergibt den Mittelwert der Ränge als

$$\overline{rg}_X = \overline{rg}_Y = \frac{(n+1)}{2}.$$

Bsp. Hilfsgrößen Rangkorr. (1)

i	$rg(x_i)$	$rg(y_i)$	$(rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)$
1	3	1	3
2	1	2	6
3	6	5.5	3
4	5	4	0
5	2	3	2
6	4	5.5	0
7	7	7	9
Σ	28	28	23
$\overline{rg}_X$	4		
$\overline{rg}_Y$		4	

Bsp. Hilfsgrößen Rangkorr. (2)

i	$rg(x_i)$	$rg(y_i)$	$(rg(x_i) - \overline{rg}_X)^2$	$(rg(y_i) - \overline{rg}_Y)^2$
1	3	1	1	9
2	1	2	9	4
3	6	5.5	4	2.25
4	5	4	1	0
5	2	3	4	1
6	4	5.5	0	2.25
7	7	7	9	9
Σ	28	28	28	27.5
$\overline{rg}_X$	4			
$\overline{rg}_Y$		4		

Rangkorrelationskoeffizient (3)

- Maßzahl für die Stärke des **monotonen Zusammenhangs** zwischen zwei Rangreihen
- $r_{\mathrm{SP}} > 0$: positiver monotoner Zusammenhang, Tendenz: x groß $\longleftrightarrow$ y groß, x klein $\longleftrightarrow$ y klein
- $r_{\mathrm{SP}} < 0$: negativer monotoner Zusammenhang, Tendenz: x groß $\longleftrightarrow$ y klein, x klein $\longleftrightarrow$ y groß
- $r_{\mathrm{SP}} = 0$: kein monotoner Zusammenhang
- $r_{\mathrm{SP}} = 1$, $\left(r_{\mathrm{SP}} = -1\right)$: alle Rangdaten auf einer Geraden mit positiver (negativer) Steigung liegend

Bsp. Rangkorrelationskoeffizient

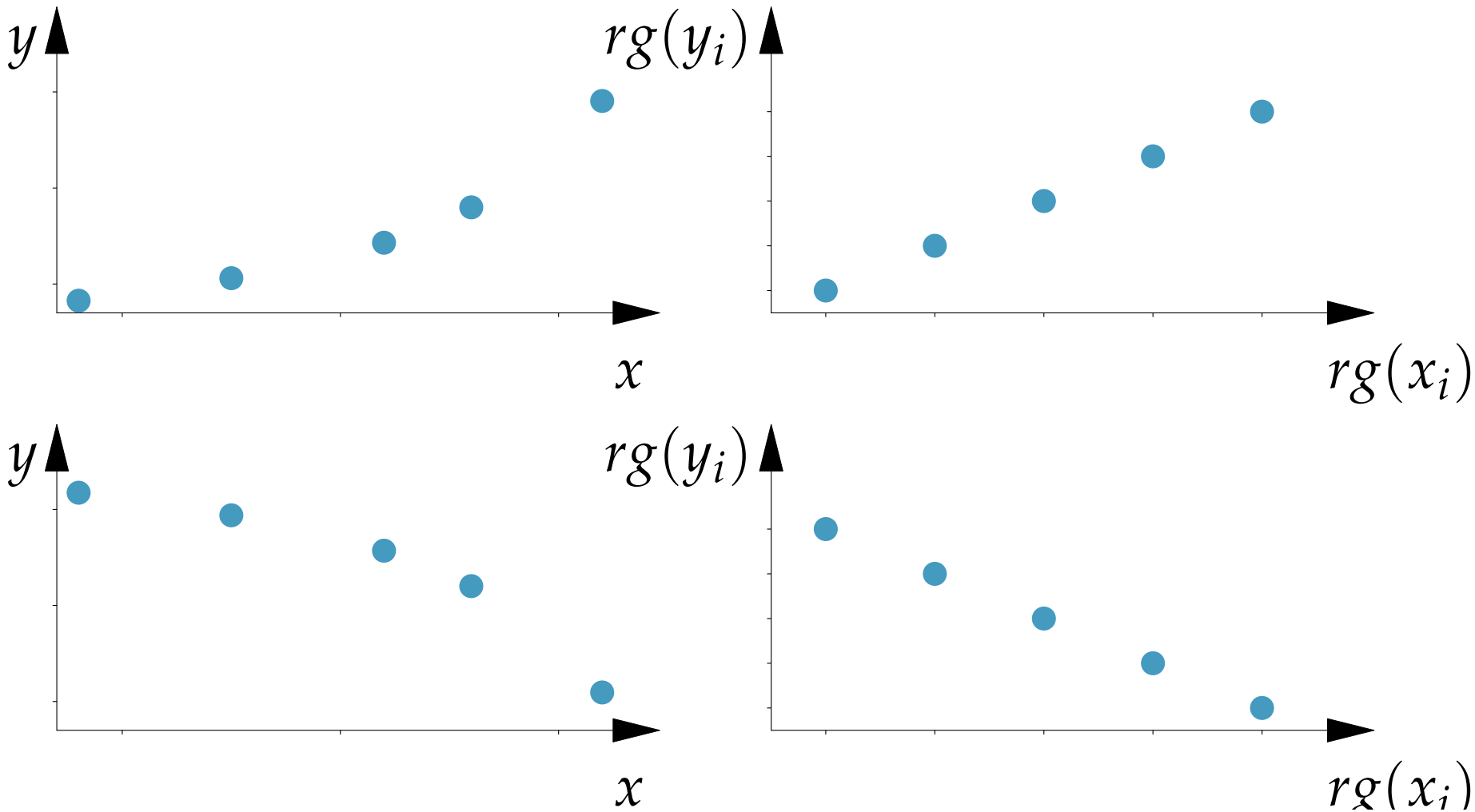


Abb.: Extremfälle für Spearmans Korrelationskoeff., $r_{\mathrm{SP}} = 1$ (oben) und $r_{\mathrm{SP}} = -1$ (unten)

Visualisierung für zwei Gruppen

- Visualisierung der Verteilung des intervall- oder verhältnisskalierten Merkmals für zwei Gruppen eines dichotomen Merkmals:
  - Streudiagramm (auch *Stripchart*)
  - Fehlerbalkendiagramm
  - Box-Plot

Bsp. Streudiagramm für zwei Gruppen

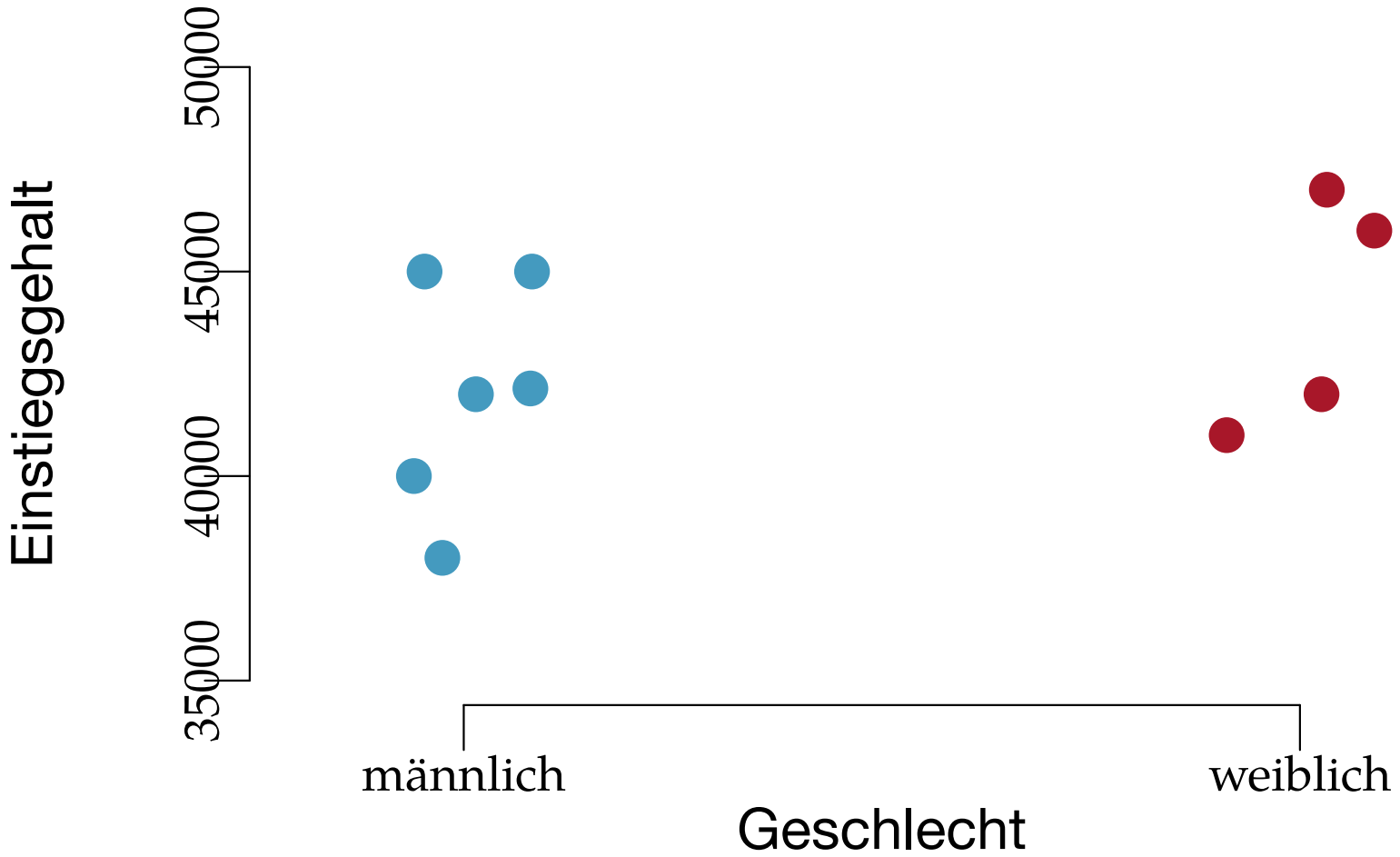


Abb.: Streudiagramm für Einstiegsgehalt und Geschlecht

Fehlerbalkendiagramm

- **Fehlerbalkendiagramm:** grafische Darstellung des arithmetischen Mittelwerts und der Standardabweichung von Daten
  - Lage des Mittelwerts wird durch einen Punkt (oder einen Balken) markiert.
  - Streuung der Werte wird durch vertikale Linien nach oben bzw. unten dargestellt.
- Einsatzgebiet: vergleichende Gegenüberstellung von Daten
- Alternative: Box-Plot (informativer, insbesondere bei kleinen Fallzahlen)

Bsp. Fehlerbalkendiagramm

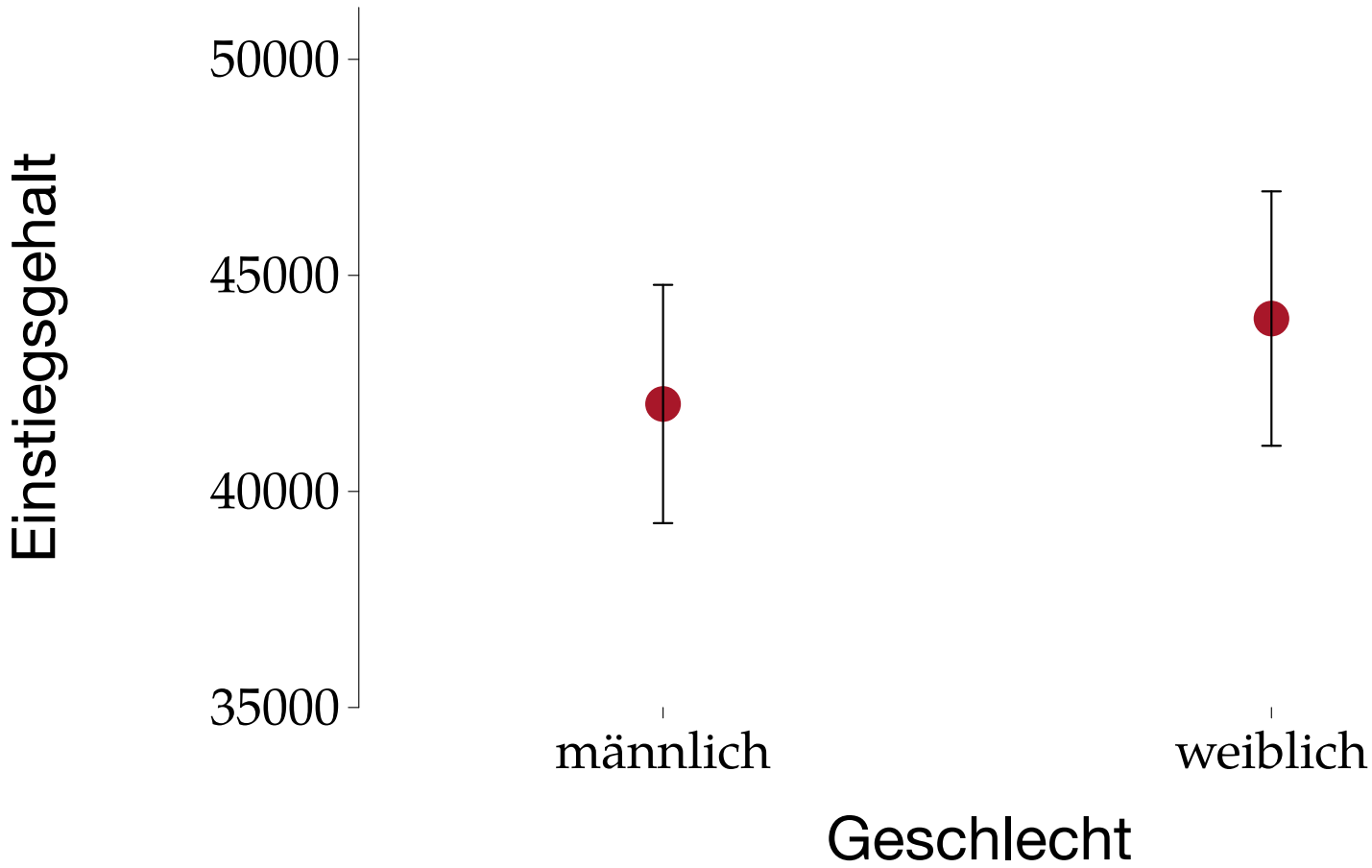


Abb.: Fehlerbalkendiagramm von Einstiegsgehalt ($\bar{x} \pm s$) für die Ausprägungen von Geschlecht

Spezialfall der Pearson-Korrelation

- Um den Zusammenhang zwischen einem dichotomen Merkmal X und einem intervall- oder verhältnisskalierten Merkmal Y zu beschreiben, kann die punktbiseriale Korrelation berechnet werden.
- Vorgehen:
  1. Die zwei Ausprägungen des dichotomen Merkmals X werden binär kodiert ($X \in \{0, 1\}$).
  2. Anschließend wird der Pearson-Korrelations-koeffizient zwischen X und Y berechnet.
- Die punktbiseriale Korrelation gibt an, wie stark sich die Mittelwerte von Y zwischen den beiden durch X definierten Gruppen unterscheiden.

Überblick Korrelationsmaße

- Zusammenhangsmaße für Häufigkeitsdaten
  - χ^2-Koeffizient
  - Kontingenzkoeffizient
  - normierter Kontingenzkoeffizient
- Zusammenhangsmaß für Ordinal-/Rangdaten oder klassierte intervall- oder verhältnisskalierte Daten
  - Spearmans r_{SP}
- Zusammenhangsmaße für intervall- oder verhältnisskalierte Daten
  - empirische Kovarianz
  - Pearsons r (oft auch als ρ bezeichnet)

Senkung des Skalenniveaus

- Sind die Skalenniveaus der Merkmale verschieden, so ist jeweils das Maß für das „niedrigere“ Skalenniveau zu verwenden (Niveauangleichung).

Bem. Korrelationsmaße

- Korrelation ist ein Maß für die *Stärke* des Zusammenhangs zwischen X und Y. Die *Richtung* der Wirkung – sofern vorhanden – wird durch Korrelationskoeffizienten nicht erfasst (also $X \leftrightarrow Y$ und nicht $X \rightarrow Y$).
- Korrelationen messen grundsätzlich nur einen *statistischen* Zusammenhang und machen keinerlei Aussage über eventuelle *kausale* Zusammenhänge!
- Korrelationsmaße für ordinal-, intervall- oder verhältnisskalierte Daten sollten immer nur im Zusammenhang mit dem zugehörigen Streudiagramm interpretiert werden.

Richtung in der Beeinflussung

- Sachlogische Überlegungen legen häufig eine Richtung in der Beeinflussung nahe, z. B.:
  - Merkmale Körpergröße, Körpergewicht:
    Die Körpergröße hat einen Einfluss auf das Körpergewicht, nicht umgekehrt.

  Damit lässt sich also ein Merkmal, sagen wir Y, als abhängig von dem anderen Merkmal X ansehen.
- Frage: Wie beeinflusst eine unabhängige Variable X die abhängige Variable Y?

Bsp. Schlafverhalten

- Für 9 Kinder (zufällig ausgewählt, gleiches Alter) wurden die Fernsehzeit X und die Dauer Y der Tiefschlafphasen einer Nacht erhoben (jeweils in Stunden):

i	1	2	3	4	5	6	7	8	9
x_i	0.3	2.2	0.5	0.7	1.0	1.8	3.0	0.2	2.3
y_i	5.8	4.4	6.5	5.8	5.6	5.0	4.8	6.0	6.1

Bsp. Streudiagramm Schlafverhalten

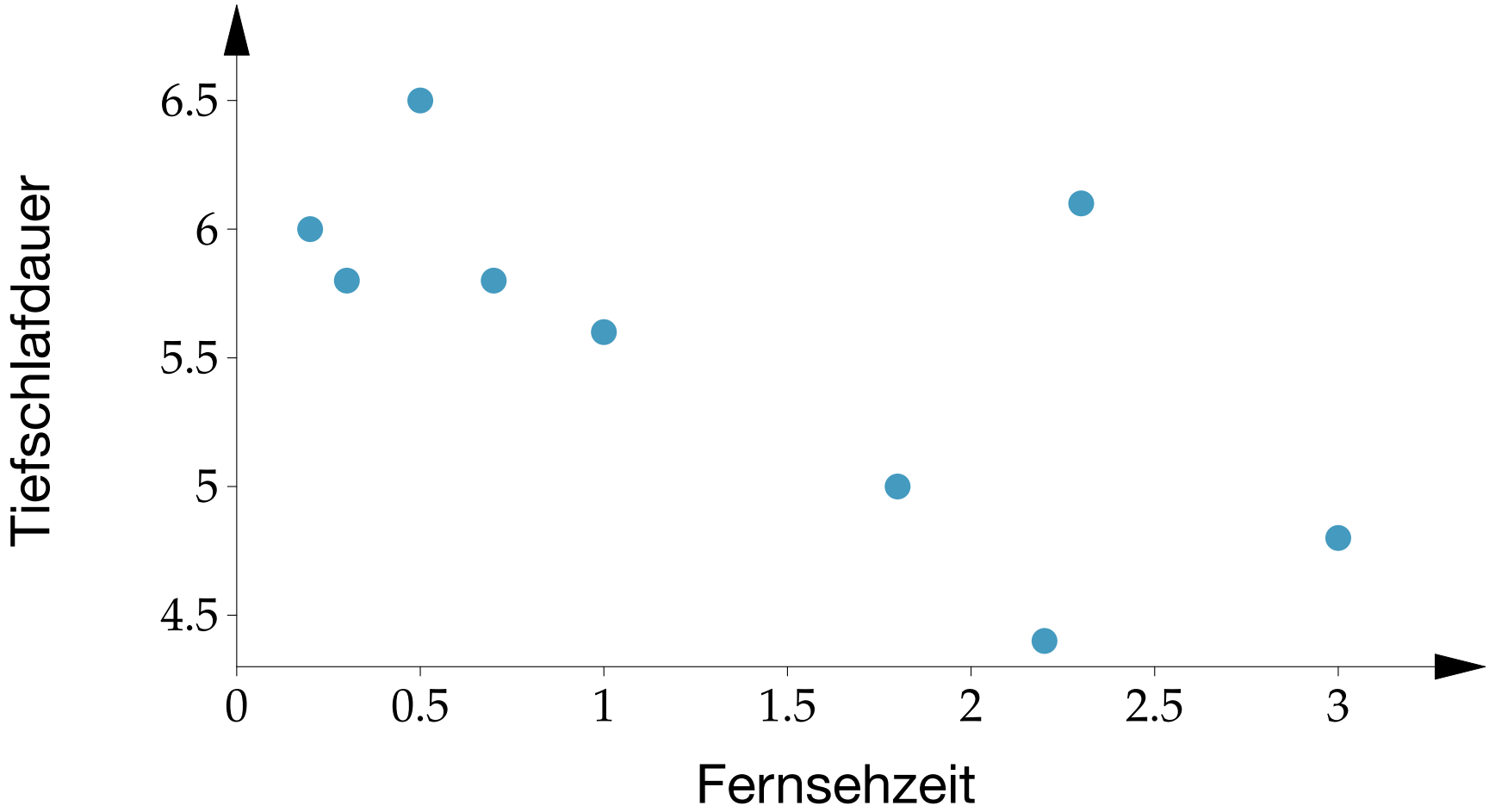


Abb.: Streudiagramm von Fernsehzeit X und Dauer Y der Tiefschlafphasen jeweils in Stunden

Einfache Regressionsanalyse (1)

- Beschreibung des Zusammenhangs zwischen zwei Merkmalen Y und X durch eine funktionale Beziehung der Form

$$Y = f(X) + \varepsilon\,.$$

  Dabei ist f eine deterministische Regressionsfunktion und ε der „Fehler“, der die Abweichung zwischen dem Modell und den Daten beschreibt.

- In dem Fall, dass eine Gerade an die Daten anpasst werden soll, ist die Funktion f somit von der Gestalt

$$f(X) = \beta_0 + \beta_1 X\,.$$

Bsp. Regressionsfkt. Schlafverhalten

- Regression von Tiefschlafdauer auf Fernsehzeit:

$$\text{Tiefschlafdauer} = \underbrace{\beta_0 + \beta_1 \cdot \text{Fernsehzeit}}_{f(\text{Fernsehzeit})} + \varepsilon$$

Einfache Regressionsanalyse (2)

- Seien $(y_1, x_1), \ldots, (y_n, x_n)$ Beobachtungen der Merkmale Y und X, dann heißt

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n\,,$$

lineare Einfachregression, wobei β_0 den Achsenabschnitt, β_1 den Steigungsparameter und ε den Fehler bezeichnen.

Berechnung der Ausgleichsgeraden

- Bei der Beschreibung der Beobachtungen durch eine Gerade $\beta_0 + \beta_1 X$ wird man versuchen, die Koeffizienten β_0 und β_1 so zu bestimmen, dass die einzelnen Datenpunkte möglichst wenig von der Gerade entfernt liegen.
- Wir bezeichnen die prognostizierten Werte als $\hat{y}_i$, $i = 1, \ldots, n$.

Bsp. Ausgleichsgeraden Schlafverh. (1)

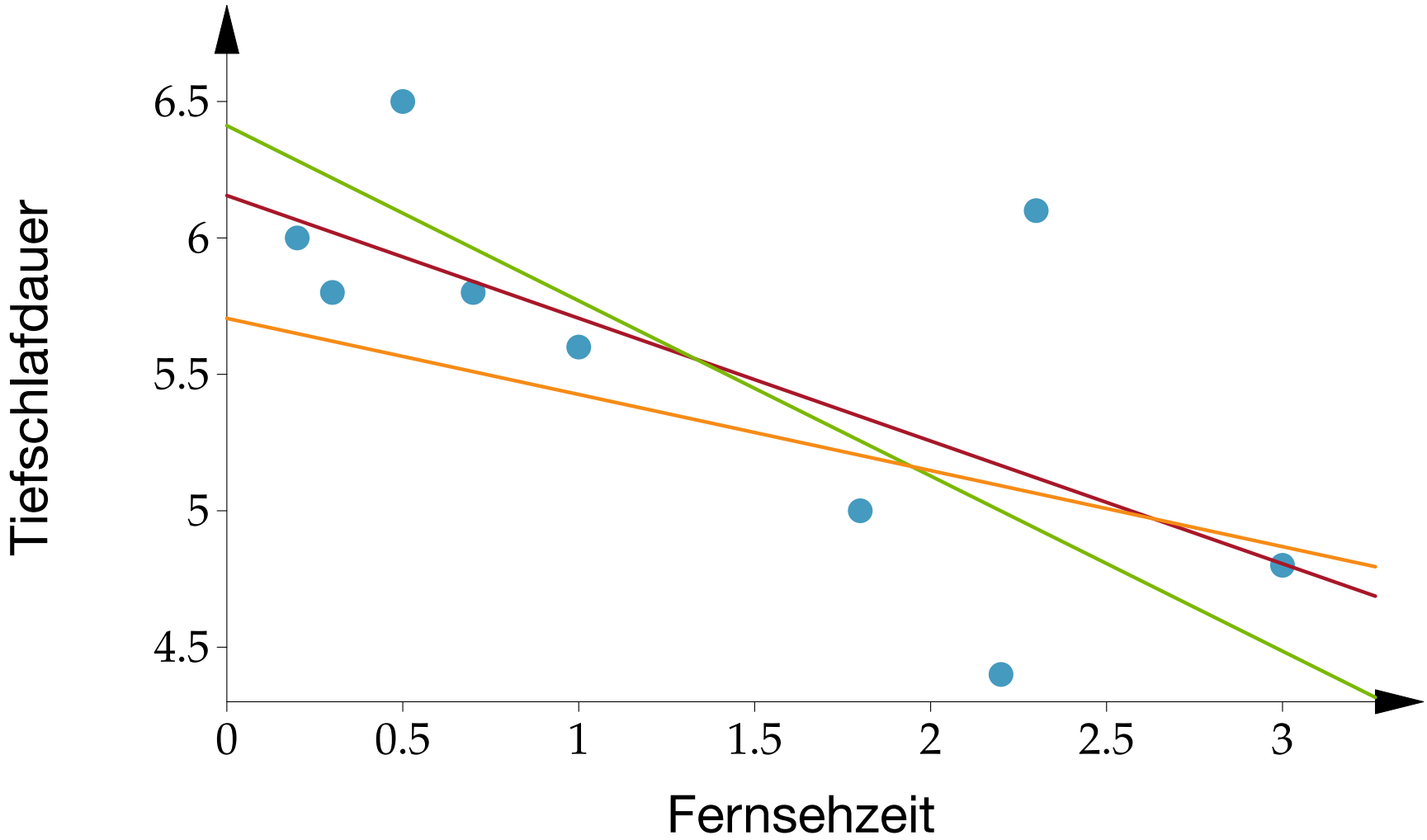
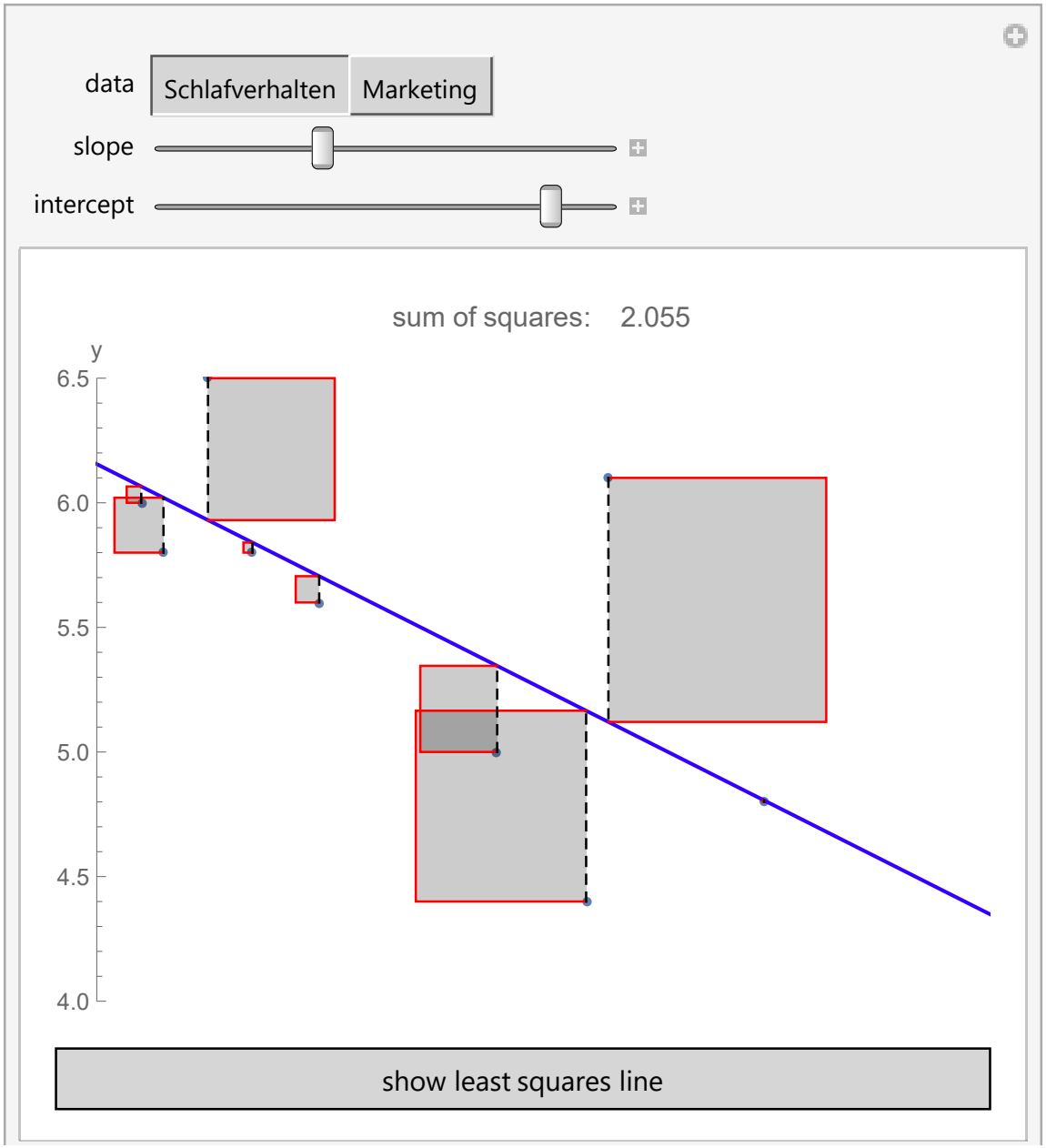


Abb.: Streudiagramm und mögliche Ausgleichsgeraden

Bsp. Ausgleichsgeraden Schlafverh. (2)

Out[ ]=



Kleinste-Quadrate-Methode

- Bestimme die Schätzungen für β_0 und β_1 so, dass die Funktion

$$Q(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

  minimiert wird.

- Die Schätzungen b_0 und b_1 für β_0 und β_1 lassen sich ermitteln, indem man $Q(\beta_0, \beta_1)$ nach β_0 und β_1 differenziert und gleich null setzt.

Kleinste-Quadrate-Schätzungen

- Als Lösung ergeben sich die **Kleinste-Quadrate-Schätzungen** b_0 und b_1 mit

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$b_0 = \bar{y} - b_1\bar{x}\,.$$

Bsp. Schlafverhalten

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	0.3	5.8	-1.03	0.24	1.07	-0.25
2	2.2	4.4	0.87	-1.16	0.75	-1
3	0.5	6.5	-0.83	0.94	0.69	-0.79
4	0.7	5.8	-0.63	0.24	0.4	-0.15
5	1	5.6	-0.33	0.04	0.11	-0.01
6	1.8	5	0.47	-0.56	0.22	-0.26
7	3	4.8	1.67	-0.76	2.78	-1.26
8	0.2	6	-1.13	0.44	1.28	-0.5
9	2.3	6.1	0.97	0.54	0.93	0.53
Σ	12	50			8.24	-3.71
$\bar{x}$	1.33					
$\bar{y}$		5.56				

Bsp. Regressionsgerade Schlafverh. (1)

- Die aus den Daten bestimmte Regressionsgerade lautet:

$$\text{Tiefschlafdauer} = \underbrace{6.16}_{b_0} - \underbrace{0.45}_{b_1} \cdot \text{Fernsehzeit}$$

- Interpretation:
  - 6.16 Stunden Tiefschlaf ohne Fernsehen
  - pro Stunde Fernsehzeit 0.45 Stunden weniger Tiefschlaf

Bsp. Regressiongerade Schlafverh. (2)

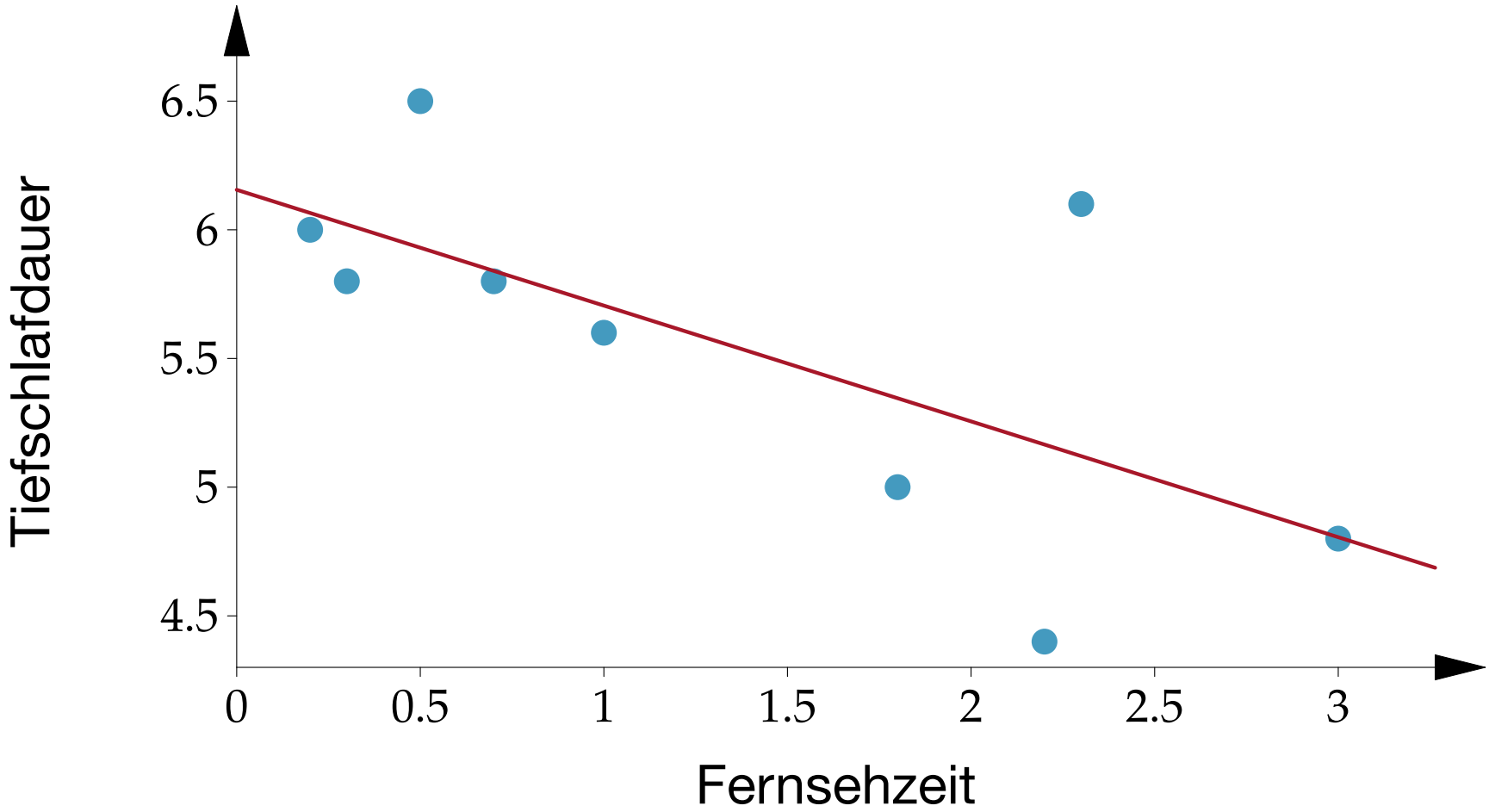


Abb.: Streudiagramm und Ausgleichsgerade zur Regression der Dauer des Tiefschlafs auf die Fernsehzeit