

Statistik

CH.4 - Zweidimensionale Verteilungen

SS 2022 | | Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

- Ziel 1
- Ziel 2
- Ziel 3

Ausgangspunkt:

- Jede statistische Einheit einer Grundgesamtheit trägt eine Vielzahl von Merkmalen.
- In diesem Kapitel werden zwei Merkmale gleichzeitig untersucht.
- Bei der Darstellung und Analyse von Abhängigkeiten zwischen Variablen muss das Skalenniveau berücksichtigt werden.

Beispiel:

- Studierende
 - ▶ Beispiel: Körpergröße und Gewicht → Streudiagramm
 - ▶ Beispiel: Geschlecht und Studiengang → Kontingenztafel
- Kraftfahrzeuge
 - ▶ Beispiel: Höchstgeschwindigkeit und Motorleistung
 - ▶ Beispiel: Kraftstoffverbrauch und Getriebeart (Manuell/Automatik)

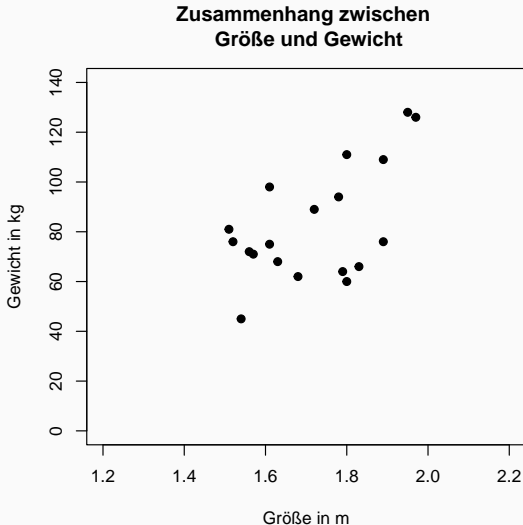
1 Streudiagramme

2 Kovarianz und Korrelation

Beispiel: Streudiagramm

Größe (m)	Gewicht (kg)
1.63	68
1.51	81
1.56	72
1.95	128
1.80	60
1.79	64
1.78	94
1.68	62
1.89	109
1.61	75
1.89	76
1.97	126
1.61	98
1.57	71
1.83	66
1.80	111
1.72	89
1.52	76
1.54	45

R-Befehl: `plot()`



1 Streudiagramme

2 Kovarianz und Korrelation

$$s_{XY} = \text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$$

- Die oben stehende Formeln gibt die Kovarianz zwischen X und Y an.
- Das Vorzeichen der Kovarianz ist ein Indikator für die Richtung eines bestehenden **linearen** Zusammenhanges zwischen Y und X.
- Die Kovarianz erlaubt es nicht Aussagen über die Stärke eines Zusammenhanges zu treffen.
- Die Größe der Kovarianz ist abhängig von der zugrundeliegenden Einheit. Einheitenwechsel (z.B. von Euro zu TEuro) führen zu einer Wertveränderung.
- **R-Befehl:** `cov()`

$$\text{Cor}(Y, X) = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i - \bar{y}}{s_y} \right) \left(\frac{x_i - \bar{x}}{s_x} \right) = \frac{\text{Cov}(Y, X)}{s_y s_x} = \frac{s_{xy}}{s_x \cdot s_y}$$

- Der Korrelationskoeffizient ist ein Maß für die Stärke des linearen Zusammenhanges.
- Im Unterschied zur Kovarianz, ist $\text{Cor}(Y, X)$ nicht skalenabhängig und erlaubt die Einschätzung von Stärke und Richtung eines linearen Zusammenhanges.
- **R-Befehl:** `cor()`

$\text{Cor}(Y, X) = 0$ bedeutet nicht, dass es zwischen X und Y keinen Zusammenhang gibt.

- 1 Wertebereich: $-1 \leq r_{XY} \leq 1$
- 2 Ist $r_{XY} = 0$, so sind X und Y nicht korreliert (unkorreliert).
- 3 Ist $r_{XY} > 0$, so sind X und Y gleichläufig (gleichsinnig) korreliert.
- 4 Ist $r_{XY} < 0$, so sind X und Y gegenläufig (ungleichsinnig) korreliert.
- 5 Je größer $|r_{XY}|$ ist, desto stärker ist die Korrelation zwischen X and Y .

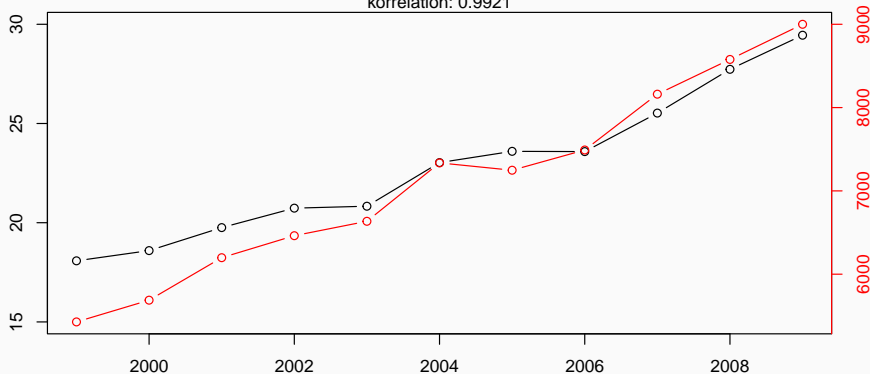
Scheinkorrelation: obwohl ein großer Wert des Korrelationskoeffizienten zwischen X und Y besteht, liegt kein *ursächlicher* (und/oder sachlogischer) Zusammenhang zwischen X und Y vor.

Beispiel

Zusammenhang zwischen Kindergeburten und der Anzahl der Storchpaare, die sich in einer Region ansiedeln.

US Spending on science, space, and technology and Suicides by hangig, strangulation and suffocation

korrelation: 0.9921



■ Weitere Beispiele unter: <http://tylervigen.com/spurious-correlations>

- Welche Darstellungsmöglichkeiten gibt es für zweidimensionale Daten?
- Bedeutet ein Korrelationskoeffizient nah bei 1, dass ein sachlicher Zusammenhang zwischen den untersuchten Merkmalen besteht?
- Wie ist ein Korrelationskoeffizient nah bei -1 zu interpretieren?