

Statistik

CH.4 - Zweidimensionale Verteilungen

2024 || Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

- Erweiterung der Streuungsbetrachtung von einem auf zwei Merkmale
- Erkennen des Zusammenhangs von Streuung mehrerer Variablen und (linearem) Zusammenhang
- Diskussion und Betrachtung relevanter Maßzahlen zur Messung von Zusammenhängen

- 1 Streuung und Streudiagramme
- 2 Kovarianz
- 3 Korrelation

Rekapitulation: Streuung



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)

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 Streuung und Streudiagramme

2 Kovarianz

3 Korrelation

Gewicht (kg) vs. Körpergröße (cm)



Aufgabe: Bestimmen des Vorzeichens

- $y_i - \bar{y}$ ist die Differenz jeder Beobachtung y_i vom arithmetischen Mittel der abhängigen Variablen
- $x_i - \bar{x}$ ist die Abweichung x_i vom arithmetischen Mittel des Prädiktors
- $(y_i - \bar{y})(x_i - \bar{x})$ ist das Produkt der vorherigen beiden Größen

| Quadrant | $y_i - \bar{y}$ | $x_i - \bar{x}$ | $(y_i - \bar{y})(x_i - \bar{x})$ |
|------------------|-----------------|-----------------|----------------------------------|
| 1 (oben rechts) | | | |
| 2 (oben links) | | | |
| 3 (unten links) | | | |
| 4 (unten rechts) | | | |

Positiver Zusammenhang

- Wenn der Zusammenhang zwischen Y und X **positiv** ist (also wenn X größer wird, dann wird auch Y größer), dann sind mehr Datenpunkte im ersten und dritten Quadranten als im zweiten und vierten.
- Die Summe der Elemente in der letzten Spalte der vorherigen Tabelle ist dann mit großer Wahrscheinlichkeit positiv, also $\text{Cov}(Y, X) > 0$.

Positiver Zusammenhang

- Wenn der Zusammenhang zwischen Y und X **positiv** ist (also wenn X größer wird, dann wird auch Y größer), dann sind mehr Datenpunkte im ersten und dritten Quadranten als im zweiten und vierten.
- Die Summe der Elemente in der letzten Spalte der vorherigen Tabelle ist dann mit großer Wahrscheinlichkeit positiv, also $\text{Cov}(Y, X) > 0$.

Negativer Zusammenhang

- Wenn der lineare Zusammenhang zwischen Y und X **negativ** ist (z.B. wenn X sinkt, steigt Y), dann befinden sich mehr Datenpunkte im zweiten und vierten Quadranten als im ersten und dritten.
- Die Summe der Elemente in der letzten Spalte der vorherigen Tabelle ist dann mit großer Wahrscheinlichkeit negativ, also $\text{Cov}(Y, X) < 0$.

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

- 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** Zusammenhangs zwischen Y und X.
- Die Kovarianz erlaubt es nicht, Aussagen über die Stärke eines Zusammenhangs 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()`

1 Streuung und Streudiagramme

2 Kovarianz

3 Korrelation

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

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