Bivariate Statistik

Empirische Kovarianz

Def.: Die empirische Kovarianz ist eine (nicht normierte) Maßzahl für den linearen Zusammenhang (Korrelation) zwischen zwei Merkmalen.

$$\tilde{s}_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \boxed{\overline{xy} - \bar{x} \cdot \bar{y}} \qquad \tilde{s}_{xx} = \tilde{s}_x^2$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{k} n_i \cdot (x_i - \bar{x}) \cdot (y_i - \bar{y}) \quad \text{gruppierte Datenpaare}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} \cdot (x_i - \bar{x}) \cdot (y_j - \bar{y}) \quad \text{gruppiert in Kontingenztabelle}$$

Anmerkung: Bei klassierten Daten Klassenmitten $\bar{x}_i$ und $\bar{y}_i$ statt x_i und y_i verwenden.

X, Y unkorreliert $\Leftrightarrow \tilde{s}_{xy} = 0$

Korrelationskoeffizient (Bravais-Pearson)

Def.: Der Korrelationskoeffizient (nach Bravais/Pearson) misst den Grad und die Richtung der linearen Abhängigkeit zwischen zwei Merkmalen.

$$r_B = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}$$

$r_B \in [-1, 1]$

- $r_B \to -1$: neg. linearer Zusammenhang
- $r_B \to 1$: pos. linearer Zusammenhang

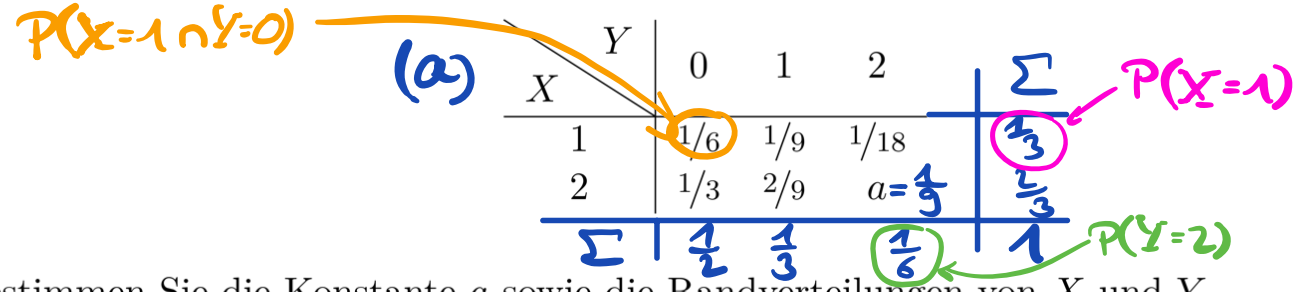
X, Y unkorreliert $\Leftrightarrow \tilde{s}_{xy} = 0 \Leftrightarrow r_B = 0$

Anmerkung:

X, Y unabhängig $\Rightarrow$ X, Y unkorreliert (nicht linear abhängig)

X, Y korreliert $\Rightarrow$ X, Y abhängig (linear abhängig)

Aufgabe: Eine Befragung junger Familien im Jahr 2023 in London über die Anzahl ihrer Kinder (X) und die Anzahl ihrer Autos (Y) ergab folgende Korrelationstabelle relativer Häufigkeiten:



(a) Bestimmen Sie die Konstante a sowie die Randverteilungen von X und Y.

(b) Sind die Merkmale X und Y unabhängig? Sind sie korreliert? Bestimmen Sie die Kovarianz.

(c) Mit welcher Wahrscheinlichkeit hat eine zufällig ausgewählte Familie

(i) zwei Kinder und mindestens ein Auto?

(ii) mit zwei Kindern mindestens ein Auto?

(iii) zwei Kinder oder mindestens ein Auto?

(a) X, Y unabh. $\Leftrightarrow P(X=x_i \cap Y=y_j) = P(X=x_i) \cdot P(Y=y_j) \quad \forall i,j$

$\frac{1}{6} = \frac{1}{3} \cdot \frac{1}{2}$ ✓

$\frac{1}{9} = \frac{1}{3} \cdot \frac{1}{3}$ ✓

$\vdots$

$\frac{1}{9} = \frac{2}{3} \cdot \frac{1}{6}$ ✓

$\Rightarrow X, Y$ unabh.

(a) $\Rightarrow X, Y$ unkorreliert

$\Rightarrow \tilde{s}_{xy} = 0$

$P(X=1 \cap Y=0)$

(a)

X \ Y	0	1	2	Σ
1	1/6	1/9	1/18	1/3
2	1/3	2/9	$a=\frac{1}{9}$	2/3
Σ	1/2	1/3	1/6	1

$P(X=1)$

$P(Y=2)$

(c) Mit welcher Wahrscheinlichkeit hat eine zufällig ausgewählte Familie

(i) zwei Kinder und mindestens ein Auto?

(ii) mit zwei Kindern mindestens ein Auto?

(iii) zwei Kinder oder mindestens ein Auto?

(i) $P(X=2 \cap Y \geq 1) = P(X=2 \cap Y=1) + P(X=2 \cap Y=2) = \frac{2}{9} + \frac{1}{9} = \frac{1}{3}$

(ii) $P(Y \geq 1 \mid X=2) = \frac{P(X=2 \cap Y \geq 1)}{P(X=2)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$

(iii) $P(X=2 \cup Y \geq 1) = P(X=2) + P(Y \geq 1) - P(X=2 \cap Y \geq 1)$

$= \frac{2}{3} + \frac{1}{3} + \frac{1}{6} - \frac{1}{3}$

$= \frac{5}{6}$

Aufgabe: Untersucht werden soll die Korrelation zwischen Körpergewicht X in kg und Nüchternblutzucker Y in mmol/L bei einer Gruppe von zehn Erwachsenen innerhalb einer klinischen Studie. Die Daten repäsentieren Messungen, die nach einer Fastenperiode von mindestens 8 Stunden durchgeführt wurden, um die natürliche Regulierung des Blutzuckers durch den Körper zu erfassen.

x_i	62	68	71	73	78	82	85	89	90	94
y_i	4,9	5,0	5,5	5,2	5,7	6,1	6,6	6,2	6,3	6,8

Berechnen Sie den Korrelationskoeffizienten nach Bravais-Pearson.

$$r_B = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{\overline{x^2}-\overline{x}^2}\cdot\sqrt{\overline{y^2}-\overline{y}^2}} \approx 0{,}9484$$

→ pos. Zsh.

→ „starker" linearer Zsh.

$$\overline{x} = \frac{1}{n}\cdot\sum_{i\in J} x_i = \frac{1}{10}\cdot(62+68+71+\ldots+94) = 79{,}2$$

$$\overline{y} = \frac{1}{n}\cdot\sum_{i\in J} y_i = \frac{1}{10}\cdot(4{,}9+5{,}0+\ldots+6{,}8) = 5{,}83$$

$$\overline{x^2} = \frac{1}{n}\cdot\sum_{i\in J} x_i^2 = \frac{1}{10}\cdot(62^2+68^2+\ldots+94^2) = 6372{,}8$$

$$\overline{y^2} = \frac{1}{n}\cdot\sum_{i\in J} y_i^2 = \frac{1}{10}\cdot(4{,}9^2+5{,}0^2+\ldots+6{,}8^2) = 34{,}383$$

$$\overline{xy} = \frac{1}{n}\cdot\sum_{i\in J} x_i\cdot y_i = \frac{1}{10}\cdot(62\cdot 4{,}9+\ldots+94\cdot 6{,}8) = 467{,}77$$

Korrelationskoeffizient (Spearman)

$rg(\cdot)$ = Ränge der (sortierten) Stichprobe

Def.: Der Rangkorrelationskoeffizient (nach Spearman) misst den Grad und die Richtung des **monotonen** Zusammenhangs zwischen zwei Merkmalen als Korrelation der Ränge.

$$r_S = \frac{\tilde{s}_{rg(x)\,rg(y)}}{\tilde{s}_{rg(x)} \cdot \tilde{s}_{rg(y)}} = \frac{\frac{1}{2}\left(\tilde{s}^2_{rg(x)} + \tilde{s}^2_{rg(y)}\right) - \frac{1}{2n}\sum_{i=1}^{n} d_i^2}{\tilde{s}_{rg(x)} \cdot \tilde{s}_{rg(y)}} \quad \text{mit} \quad d_i = rg(x_i) - rg(y_i)$$

Bei Bindungen: → Durchschnittsränge

- **ohne Bindungen:** $\tilde{s}^2_{rg(x)} = \tilde{s}^2_{rg(y)} = \frac{n^2-1}{12}$

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2-1)}$$

- **mit Bindungen:** (mehrdeutige Ränge)

$$\tilde{s}^2_{rg(\cdot)} = \frac{n^2-1}{12} - \sum_{j=1}^{k_\cdot} t_j \cdot \frac{t_j^2-1}{12}$$

$$r_S = \frac{n^3 - n - \frac{1}{2}(T_x^* + T_y^*) - 6 \cdot \sum_{i=1}^{n} d_i^2}{\sqrt{n^3 - n - T_x^*} \cdot \sqrt{n^3 - n - T_y^*}} \quad \text{mit} \quad T_\cdot^* = \sum_{j=1}^{k_\cdot} t_j \cdot (t_j^2 - 1)$$

t_j: Anzahl der Beobachtungen mit gleichem Rang in Klasse j

Streuungszerlegung nach Klassenbildung: $\tilde{s}^2_{Gesamt} = \tilde{s}^2_{innerhalb} + \tilde{s}^2_{zwischen}$

Aufgabe: Eine Variante des Spiels Schach, die sich der ehemalige Weltmeister Bobby Fischer ausgedacht hat, wird Fischer Random Schach oder auch Schach960 genannt. An der zweiten Weltmeisterschaft, die vom 25.-30. Oktober 2022 in Reykjavík stattfand, traten einige der stärksten Schachspieler der Welt gegeneinander an. Die Stärke von Spielern und damit ihrer Einordnung in der Weltrangliste wird durch eine Wertzahl (ELO) ausgedrückt. Wegen der unvorhersehbaren Ausgangsposition beim Fischer Random Schach (es gibt 960 verschiedene) ist eine Untersuchung des Zusammenhangs von Spielerstärke und Turnierplatzierung von Interesse.

Spieler	ELO (Rapid Rating)	rg(x)	Turnierplatzierung	rg(y)	|d_i|
Magnus Carlsen	2834	1	3	2	2
Hikaru Nakamura	2789	2	1	1	1
Wesley So	2784	3	6	3	3
Ian Nepomniachtchi	2766	4	2	2	2
Vladimir Fedoseev	2741	5	5	5	0
Nodirbek Abdusattorov	2676	6	4	4	2
Matthias Blübaum	2617	7	7	7	0
Hjörvar Steinn Grétarsson	2520	8	8	8	0

Untersuchen Sie den Zusammenhang zwischen Spielstärke und Turnierplatzierung durch einen geeigneten Korrelationskoeffizienten. Was sagt dieser aus?

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot (2^2 + 1^2 + 3^2 + 2^2 + 0^2 + 2^2 + 0^2 + 0^2)}{8 \cdot (8^2 - 1)}$$

$$= \frac{31}{42} \approx 0{,}7381 \quad (\hat{=} 73{,}81\%)$$

Quadratische Kontingenz

Def.: Die quadratische Kontingenz ist die Summe aller Verhältnisse von den quadratischen Abständen der beobachteten und unter Unabhängigkeit erwarteten Häufigkeiten zu den unter Unabhängigkeit erwarteten Häufigkeiten.

$\sim B(n, p_{ij})$ — $\mathbb{E}(n_{ij})$ unter Unabh.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}} = n \cdot \left(\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{n_{i\cdot} \cdot n_{\cdot j}} - 1 \right)$$

$\approx \mathrm{Var}(n_{ij})$ für große n (und damit kleine p_{ij})

- $\chi^2 \in [0, n \cdot (\min(k,l) - 1)]$ (Unabh. ↑, → exakter Zusammenhang)
- Kontingenzmaß nach Cramer: $V = \sqrt{\frac{\chi^2}{n \cdot (\min(k,l) - 1)}} \in [0,1]$
- korrigierter Kontingenzkoeffizient:

$$C_{korr.} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \sqrt{\frac{\min(k,l)}{\min(k,l) - 1}} \in [0,1]$$

$$\frac{n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}}{\sqrt{\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}}} \xrightarrow[n \to \infty]{\text{ZGWS}} Z \sim N(0,1)$$

$$\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}} \xrightarrow{n \to \infty} \sum_{i=1}^{k} \sum_{j=1}^{l} Z^2 \sim \chi^2_{(k-1)(l-1)}$$

Aufgabe: Am Ende des Kursjahres untersucht der Statistik Lehrstuhl einer Universität den Zusammenhang zwischen dem Besuch der Vorlesung (X) und dem Klausurergebnis (Y). Es ergab sich folgende Kontingenztabelle:

X \ Y	**Prüfung** bestehen	**Prüfung** nicht bestehen	$n_{i\cdot}$
Anwesenheit 30 – 100%	129	29	158
Anwesenheit 0 – 30%	52	30	82
$n_{\cdot j}$	181	59	$n = 240$

(a) Berechnen Sie die quadratische Kontingenz und geben Sie den Wertebereich an.

(b) Bestimmen Sie außerdem das normierte Kontingenzmaß Cramers V und den korrigierten Kontingenzkoeffizienten.

(c) Welche Aussage können Sie über den Zusammenhang zwischen den Merkmalen X und Y treffen?

(a) $$\chi^2 = n \cdot \left(\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{n_{i\cdot} \cdot n_{\cdot j}} - 1 \right) = 240 \cdot \left(\frac{129^2}{181 \cdot 158} + \frac{29^2}{59 \cdot 158} + \frac{52^2}{182 \cdot 82} + \frac{30^2}{59 \cdot 82} - 1 \right)$$

$$= 9{,}4374$$

$$\chi^2 \in [0, n \cdot (\min(k,l) - 1)] = [0, 240]$$

(b) $$V = \sqrt{\frac{\chi^2}{n \cdot (\min(k,l) - 1)}} = 0{,}1983 \qquad C_{korr.} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \sqrt{\frac{\min(k,l)}{\min(k,l) - 1}} = 0{,}2751$$

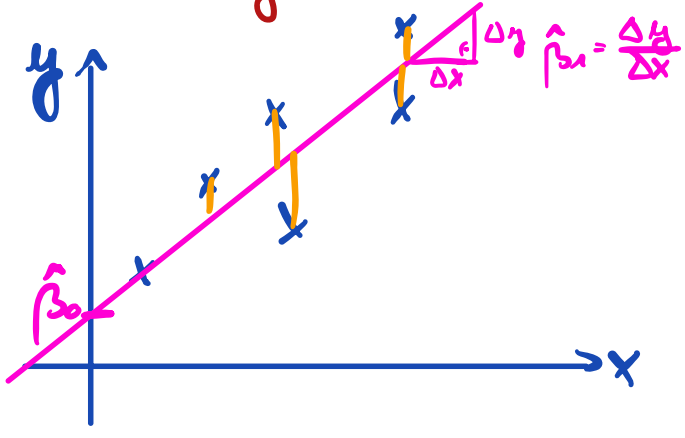
$0{,}1945$ $\quad \sqrt{2}$

Regressionsanalyse

Ziel: Erstellen eines Modells, das die abhängige Variable basierend auf den unabhängigen Variablen erklärt, um Vorhersagen zu treffen oder kausale Beziehungen zu erklären.

Idee: Suche Funktion $\hat{y} := \hat{y}(\vec{x}, \vec{\beta})$, sodass die Summe der quadratischen (vertikalen) Abweichungen von der vorliegenden Stichprobe minimal ist, also

$$Q := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \longrightarrow \min$$



Anmerkung:

- Merkmale $\vec{X}$ heißen unabhängige Variablen
- Merkmal Y heißt abhängige Variable

Lineare Regression (in einer Variable)

Def.: Für ein lineares Regressionsmodell (in einer Variable) wähle die Ausgleichsfunktion $\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x}$ (Regressionsgerade).

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} \left(= r_B \cdot \frac{\tilde{s}_y}{\tilde{s}_x}\right) \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

Eigenschaften:

- $\hat{\beta}_0$ = Schnittpunkt der Regressionsgerade mit y-Achse
- $\hat{\beta}_1$ = Anstieg der Regressionsgerade
- $\hat{\beta}_1 = 0 \iff r_B = 0 \iff X, Y$ linear unabhängig
- $(\bar{x}, \bar{y})$ liegt auf Regressionsgerade
- $\sum_{i=1}^{n} \hat{\epsilon}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$ (bzw. $\bar{\hat{y}} = \bar{y}$)

Herleitung von $\hat{\beta}_0$ und $\hat{\beta}_1$:

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)^2 \rightarrow \min$$

$$\frac{\partial Q}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i) = -2 \cdot (n \cdot \bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 \cdot n \cdot \bar{x}) \stackrel{!}{=} 0 \qquad (I)$$

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2 \cdot \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i) \cdot x_i = -2 \cdot \sum_{i=1}^{n} (x_i y_i - \hat{\beta}_0 \cdot x_i - \hat{\beta}_1 \cdot x_i^2)$$

$$= -2 \cdot (n \cdot \overline{xy} - \hat{\beta}_0 n \bar{x} - \hat{\beta}_1 n \cdot \overline{x^2}) = 0 \qquad (II)$$

aus (I): $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$

aus (II): $\overline{xy} - \hat{\beta}_0 \bar{x} - \hat{\beta}_1 \overline{x^2} = 0$

$$\overline{xy} - \bar{x}\bar{y} + \underbrace{\hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \overline{x^2}}_{= -\hat{\beta}_1 \cdot (\overline{x^2} - \bar{x}^2)} = 0 \quad \Rightarrow \quad \hat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

Aufgabe: Untersucht werden soll der lineare Zusammenhang zwischen dem Lebensalter (X) und dem monatlichen Nettoeinkommen (Y) folgender $n = 10$ befragter Personen im Jahr 1995 in Bayreuth. Stellen Sie für folgende Daten die Gleichung der Regressionsgerade auf.

x_i	22	28	32	36	40	44	48	52	56	62
y_i	1200	2400	1400	2600	1800	2800	3200	1600	3000	2000

$n = 10$

$$\sum_{i=1}^{n} x_i = 420 \quad \sum_{i=1}^{n} x_i^2 = 19112 \quad \sum_{i=1}^{n} y_i = 22000 \quad \sum_{i=1}^{n} y_i^2 = 528 \cdot 10^5 \quad \sum_{i=1}^{n} x_i y_i = 956000$$

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \left(= r_{xy} \cdot \frac{\tilde{s}_y}{\tilde{s}_x} \right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 42$$

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 = 1911{,}2$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 2200$$

$$\overline{y^2} = \frac{1}{n} \sum_{i=1}^{n} y_i^2 = 528 \cdot 10^4$$

$$\overline{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i y_i = 95600$$

$$\hat{\beta}_1 = \frac{95.600 - 42 \cdot 2200}{1911{,}2 - 42^2} = \frac{500}{23}$$

$$\hat{\beta}_0 = 2200 - \frac{500}{23} \cdot 42 = \frac{29.600}{23}$$

$$\boxed{\hat{y} = \frac{29.600}{23} + \frac{500}{23} \cdot x}$$