LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR KOMMUNIKATIONSWISSENSCHAFT UND MEDIENFORSCHUNG

BA KW | Vorlesung

Einführung in die Statistik

Korrelation

Prof. Thomas Hanitzsch

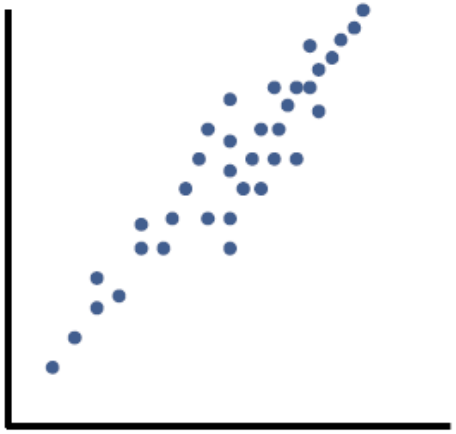
Korrelation:

Eine Beziehung zwischen zwei metrischen Variablen

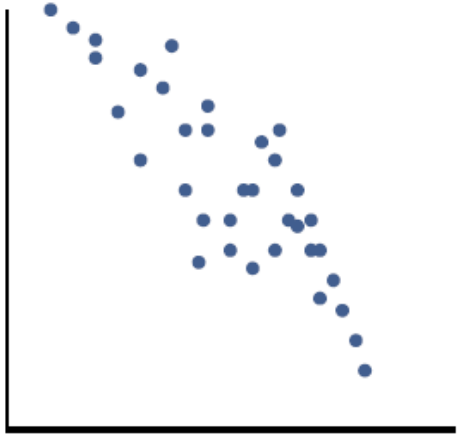
- von „Ko-Relation“ → „Wechselbeziehung“
- Ziel:
  - Herausfinden, ob und wie metrische Merkmale zusammenhängen (Stärke und Richtung des Zusammenhangs)
- Mögliche Ergebnisse:
  - **Positiver Zusammenhang** ($r_{XY} > 0$):
    je größer X, desto größer Y bzw. je kleiner X, desto kleiner Y
  - **Kein Zusammenhang** ($r_{XY} \approx 0$)
  - **Negativer Zusammenhang** ($r_{XY} < 0$):
    je größer X, desto kleiner Y bzw. je kleiner X, desto größer Y

Grafische Darstellung: das Streudiagramm

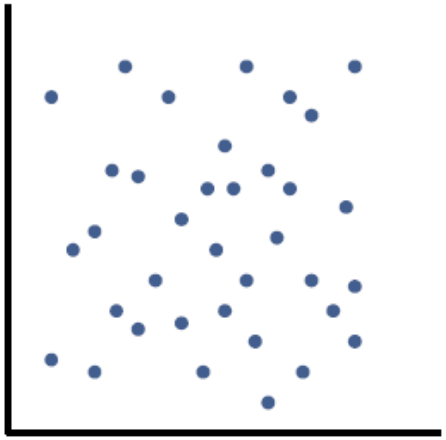
- Englisch: *scatter plot*
- Jedes Wertepaar eines Falls wird in einem Koordinatensystem als Punkt mit den Koordinaten $(x; y)$ darstellt
- Formen von Zusammenhängen:



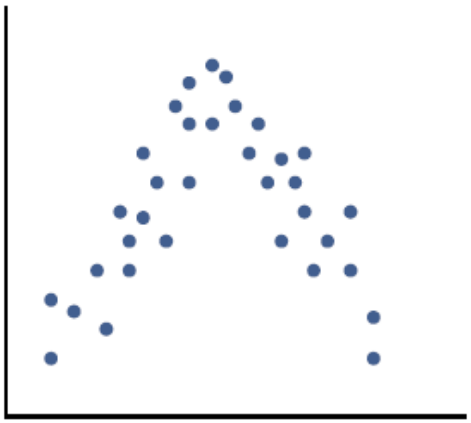
positiv linear



negativ linear



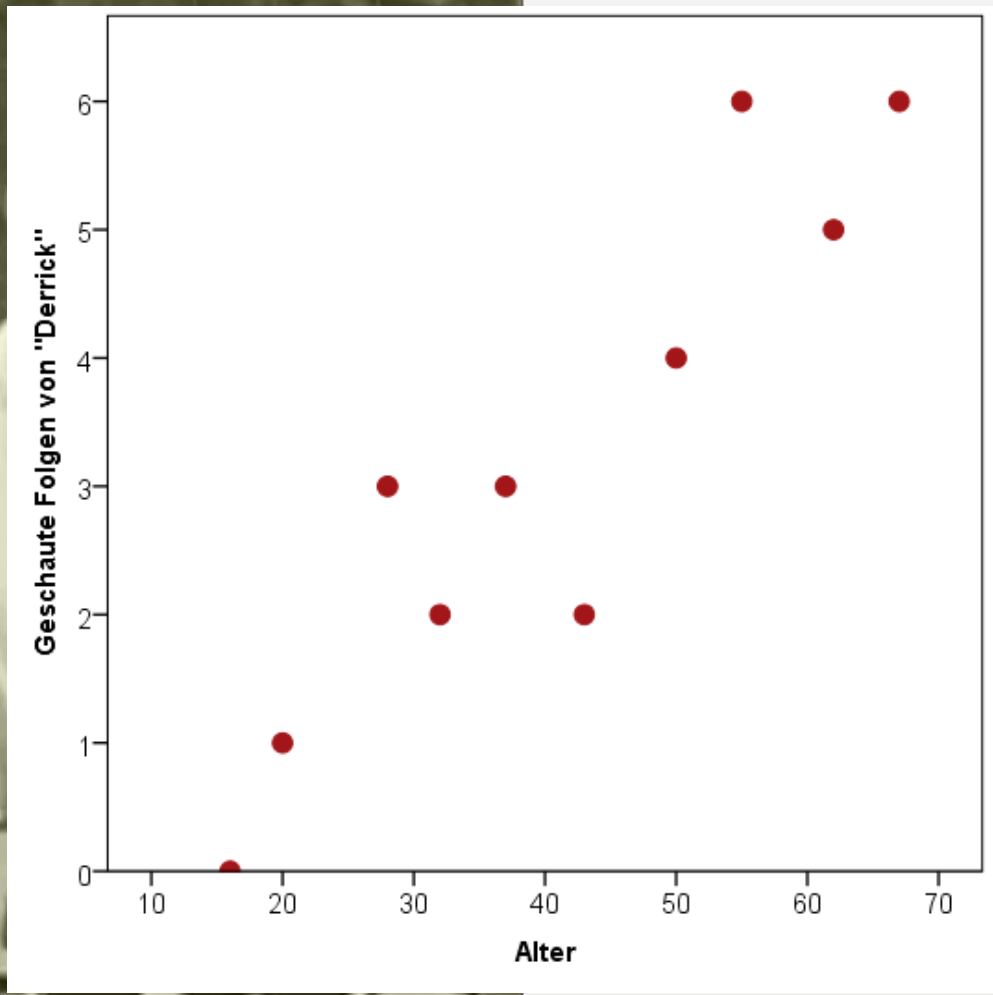
kein Zusammenhang



nicht linearer Zusammenhang (z.B. umgekehrt U-förmig)

Das Streudiagramm: Beispiel

Proband	Alter (X)	Geschaute „Derrick“-Folgen (Y)
1	62	5
2	20	1
3	28	3
4	55	6
5	50	4
6	37	3
7	43	2
8	16	0
9	32	2
10	67	6
	$\bar{x} = 41$	$\bar{y} = 3{,}2$
	$s_X = 17{,}42$	$s_Y = 2{,}04$



Die Kovarianz (s_{XY})

- Ist eine nichtstandardisierte Maßzahl für den (linearen) Zusammenhang zweier Variablen X und Y
- Wird berechnet als wechselseitige Varianz zwischen diesen Variablen:

$$s_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Die Kovarianz: Beispiel

Alter (x_i)	Geschaute Folgen von „Derrick“ (y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
62	5	62 – 41 = 21	5 – 3,2 = 1,8	21 · 1,8 = 37,8
20	1	-21	-2,2	46,2
28	3	-13	-0,2	2,6
55	6	14	2,8	39,2
50	4	9	0,8	7,2
37	3	-4	-0,2	0,8
43	2	2	-1,2	-2,4
16	0	-25	-3,2	80
32	2	-9	-1,2	10,8
67	6	26	2,8	72,8
$\bar{x} = 41$	$\bar{y} = 3{,}2$			$\Sigma = 295$

$$s_{XY} = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{10-1} \cdot 295 = \frac{295}{9} = 32{,}78$$

Die Kovarianz

- Bedeutung:
  - $s_{XY} > 0$: positive lineare Beziehung
  - $s_{XY} < 0$: negative lineare Beziehung
  - $s_{XY} \approx 0$: kein Zusammenhang
- Nachteil der Kovarianz:
  - Sie ist abhängig vom Maßstab bzw. dem Wertebereich der beobachteten Merkmale
  - Kovarianzen lassen sich daher nicht ohne Weiteres miteinander vergleichen

→ Normierung durch Korrelationskoeffizienten

Die Produkt-Moment-Korrelation (r_{XY})

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$$

wobei:

$$s_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$s_X = \sqrt{s_X^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$s_Y = \sqrt{s_Y^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

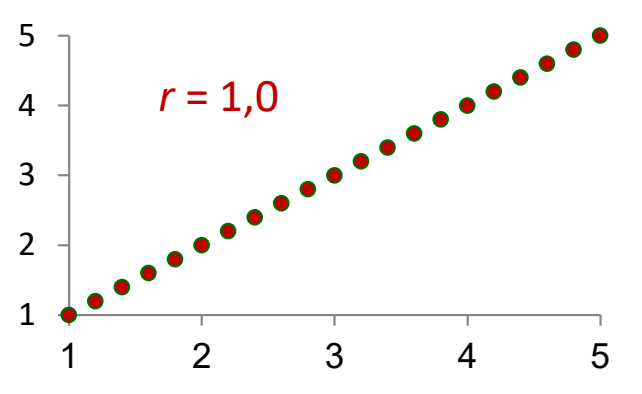
- Nach Auguste Bravais und Karl Pearson
  → „Pearson-Korrelation“ bzw. „Pearson‘s r“
- Wertebereich: von –1 bis +1
- Das Vorzeichen von r gibt die Richtung des Zusammenhangs und die Höhe von r die Stärke des Zusammenhangs wieder

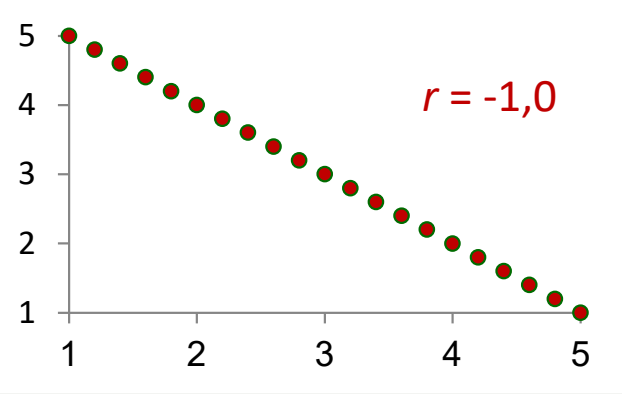
Korrelationskoeffizient: Deutung

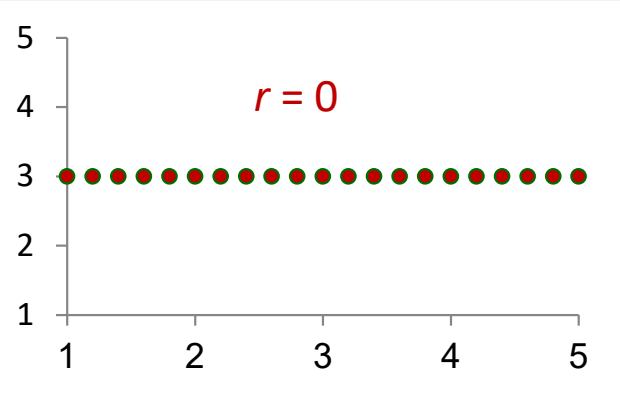
- Interpretation:
  - 1 = perfekt positiver linearer Zusammenhang
  - 0 = kein linearer Zusammenhang
  - -1 = perfekt negativer linearer Zusammenhang
- Konventionen:

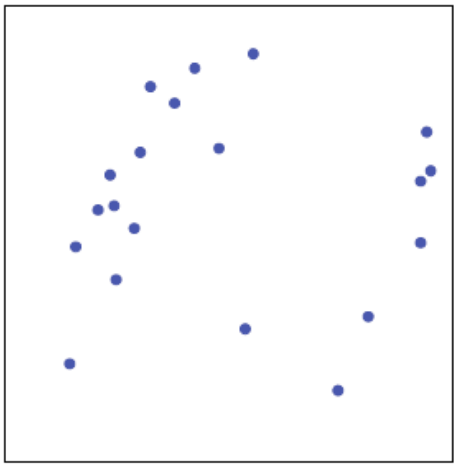
r_{XY}	Stärke des Zusammenhangs
$0{,}10 \leq \lvert r \rvert < 0{,}30$	schwacher Zusammenhang
$0{,}30 \leq \lvert r \rvert < 0{,}50$	mittlerer Zusammenhang
$0{,}50 \leq \lvert r \rvert < 0{,}70$	starker Zusammenhang
$\lvert r \rvert \geq 0{,}70$	sehr starker Zusammenhang

Korrelationskoeffizient: Deutung

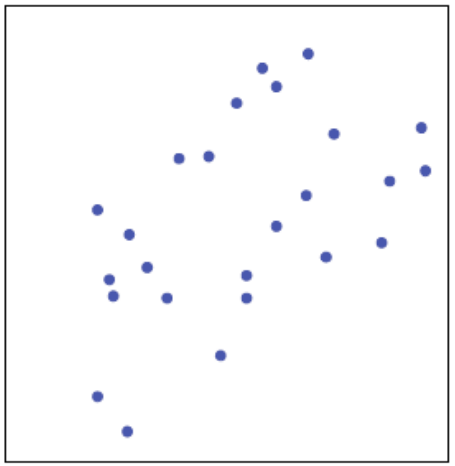




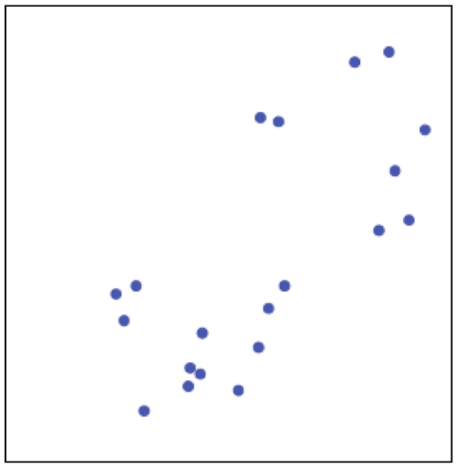




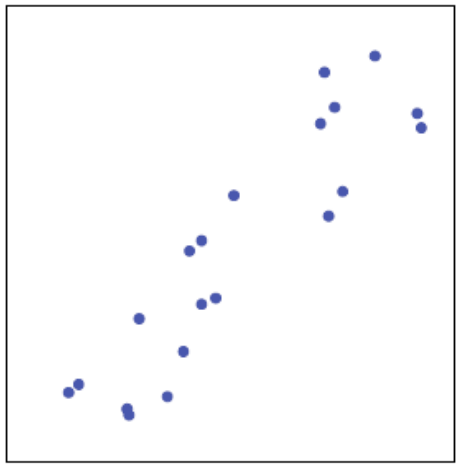
r = 0



r = 0,5

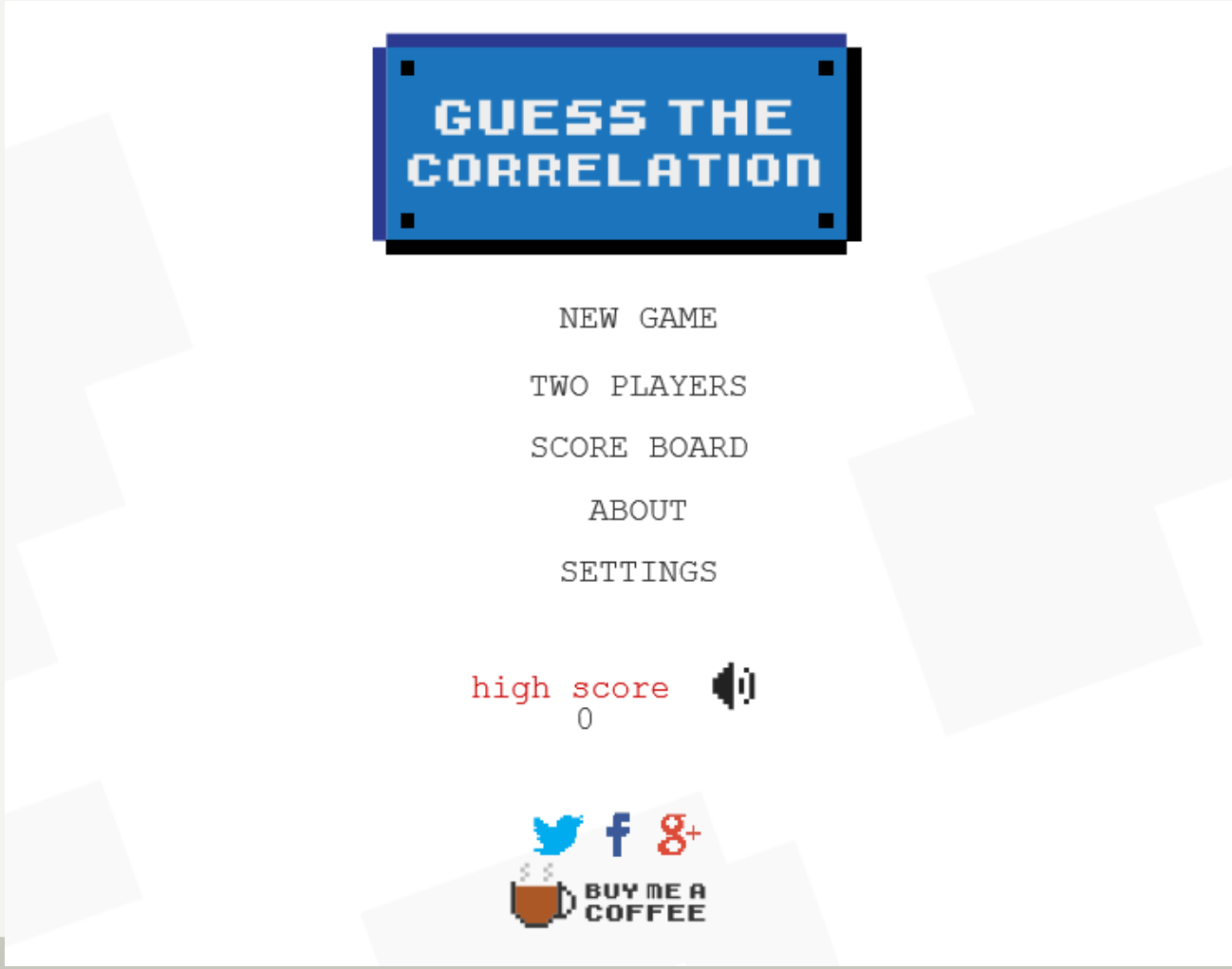


r = 0,7



r = 0,9

Korrelationskoeffizient: Deutung



Korrelationskoeffizient r_{XY} : Signifikanztest

- Funktion:
  - Prüft, ob die in der Stichprobe beobachtete Korrelation aus einer Grundgesamtheit stammt, in der in Wahrheit keine Korrelation vorliegt $(\rho_{XY} = 0)$
- Formulierung der Hypothesen:
  - Zweiseitiges Problem:

$H_0: \rho_{XY} = 0$

$H_1: \rho_{XY} \neq 0$

Korrelationskoeffizient r_{XY}: Signifikanztest

$$t = \frac{r_{XY} \cdot \sqrt{N-2}}{\sqrt{1-r_{XY}^2}} \quad \text{mit} \quad df = N - 2$$

- Testentscheidung nach festgelegtem Signifikanzniveau:
  - kritischer Wert t_{krit} kann aus der t-Tabelle ausgelesen werden (unter Berücksichtigung von α und df)
  - H_0 wird abgelehnt, wenn $|t| > t_{krit}$

Korrelationskoeffizient r_{XY}: Beispiel

Proband	Alter (X)	Geschaute „Derrick“-Folgen (Y)
1	62	5
2	20	1
3	28	3
4	55	6
5	50	4
6	37	3
7	43	2
8	16	0
9	32	2
10	67	6
	$\bar{x} = 41$	$\bar{y} = 3{,}2$
	$s_X = 17{,}42$	$s_Y = 2{,}04$

Berechnung von r_{XY}:

$$s_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{10-1} \cdot 295 = \frac{295}{9} = 32{,}78$$

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{32{,}78}{17{,}42 \cdot 2{,}04} = \mathbf{0{,}92}$$

→ sehr starker, positiver Zusammenhang:
Je älter die Menschen sind, um so mehr „Derrick“-Folgen haben sie gesehen

Korrelationskoeffizient r_{XY}: Beispiel

- Signifikanztest:

$$t = \frac{r_{XY} \cdot \sqrt{N-2}}{\sqrt{1-r_{XY}^2}} = \frac{0{,}92 \cdot \sqrt{10-2}}{\sqrt{1-0{,}92^2}} = \frac{2{,}60}{0{,}39} = \mathbf{6{,}67}$$

$$df = 10 - 2 = \mathbf{8}$$

  - **Kritischer Wert** bei $\alpha = 0{,}05$ (zweiseitig):
    - für $df = 8$: $t_{\text{krit}} = 2{,}31$
  - **Testentscheidung**:
    - $t = 6{,}67 > t_{\text{krit}}$ → H_0 wird abgelehnt

Exkurs: Arten von Korrelationen

- Rangkorrelation:
  - Zusammenhang zwischen ordinalskalierten Merkmalen (z.B. Spearmans *Rho*, ρ bzw. r_s oder Kendalls *Tau*, τ)
- Punktbiseriale Korrelation:
  - Zusammenhang zwischen einer metrischen und einer dichotomen Variable (z.B. einer 0/1-codierten „Dummy“-Variable)
- Partialkorrelation:
  - Zusammenhang zwischen zwei Variablen nach „Kontrolle“ einer oder mehrerer Drittvariablen

Korrelation und Kausalität

- Der Korrelationskoeffizient r_{XY} beschreibt den Zusammenhang nur für beobachtete Daten (d.h. die Stichprobe)
- Drittvariablenproblem: Zusammenhang könnte durch eine andere Variable begründet sein (→ „Scheinkorrelation“)
- Er gibt keinen Aufschluss über die Kausalrichtung eines Zusammenhangs
- Mögliche kausale Interpretation für eine Korrelation zwischen X und Y:
  - X beeinflusst Y
  - Y beeinflusst X
  - X und Y beeinflussen sich gegenseitig
  - Eine Drittvariable Z beeinflusst X und Y

Partieller Korrelationskoeffizient

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ} \cdot r_{YZ}}{\sqrt{1 - r^2_{XZ}} \cdot \sqrt{1 - r^2_{YZ}}}$$

- Beschreibt den Zusammenhang zwischen den Variablen X und Y, nachdem für den Einfluss der Drittvariablen Z kontrolliert wird (d.h. der Einfluss der Drittvariable wird „heraus gerechnet“)

Partieller Korrelationskoeffizient: Beispiel

Carina,

die Tasche

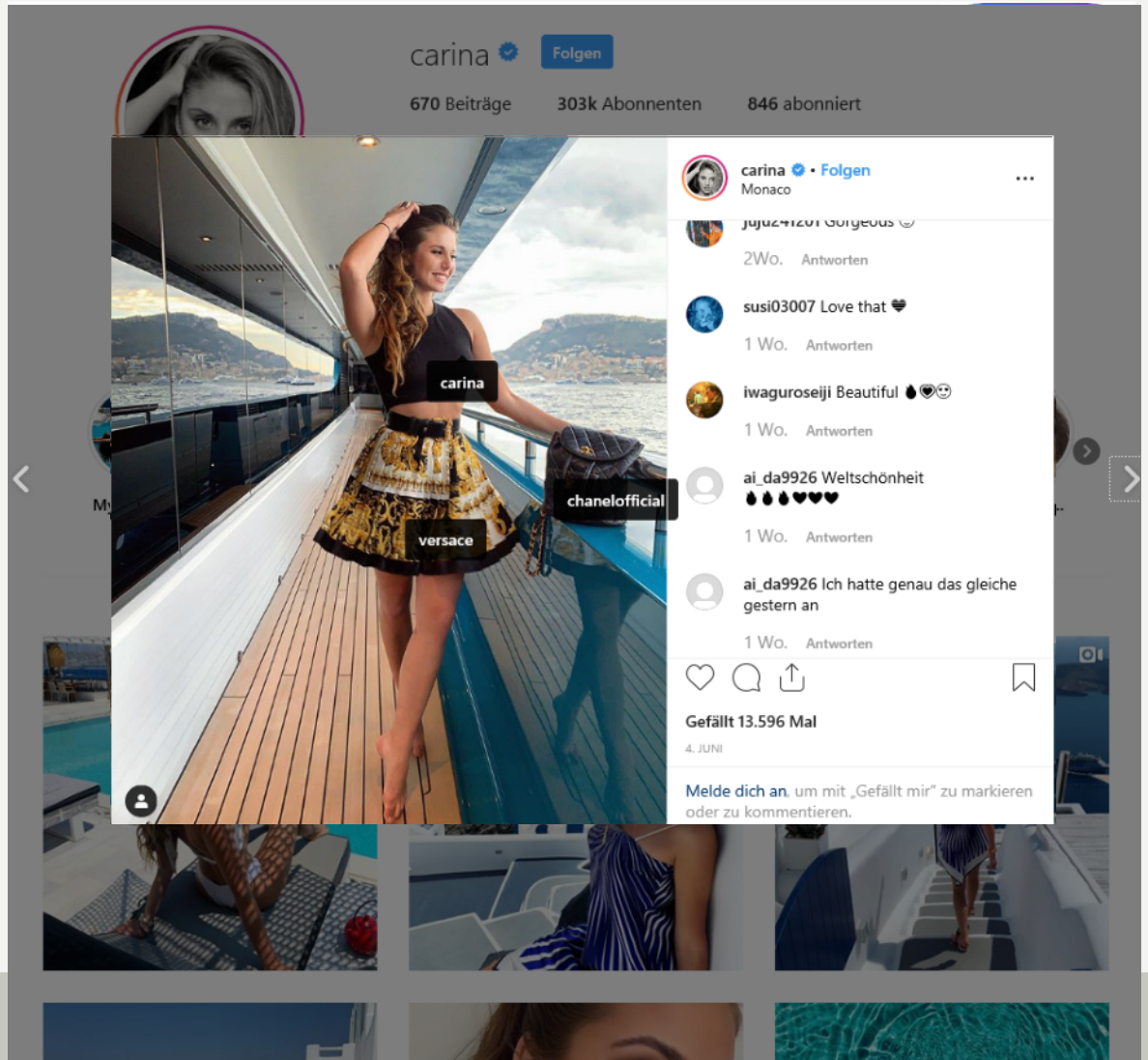
und die Kohle…

Partieller Korrelationskoeffizient: Beispiel

Carina,

die Tasche

und die Kohle…



Partieller Korrelationskoeffizient: Beispiel

Carina, die Tasche und die Kohle…

Person	1	2	3	4	5	$\bar{x}$	s
X # Werbung gesehen	5	5	15	30	25	16	11,40
Y Gekaufte Taschen	0	1	1	2	3	1,4	1,14
Z Einkommen	1500	2000	2000	3000	5000	2700	1396,42

$$r_{XY} = 0{,}827 \quad | \quad r_{XZ} = 0{,}730 \quad | \quad r_{YZ} = 0{,}958$$

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ} \cdot r_{YZ}}{\sqrt{1 - r^2{}_{XZ}} \cdot \sqrt{1 - r_{YZ}^2}} = \frac{0{,}827 - 0{,}730 \cdot 0{,}958}{\sqrt{1 - 0{,}730^2} \cdot \sqrt{1 - 0{,}958^2}} = \mathbf{0,65}$$

Aufgabe 1

Pierre erhebt bei drei zufällig ausgewählten Mitgliedern seiner Familie ($N = 3$) jeweils die tägliche Fernsehnutzungsdauer (in Stunden; $\bar{x} = 2{,}0$; $s^2 = 1{,}0$) und ihr Vertrauen in andere Menschen (Skala 1-10;$\bar{x} = 5{,}0$; $s^2 = 1{,}0$). Er nimmt entsprechend der Kultivierungshypothese an, dass TV-Nutzungsdauer und interpersonales Vertrauen negativ zusammenhängen. Er erhebt folgende Daten:

ID	TV-Nutzung	Vertrauen
1	1	5
2	2	6
3	3	4

a) Pierre setzt voraus, dass die Variablen linear zusammenhängen. Berechnen Sie ein geeignetes Maß, das einen möglichen Zusammenhang der Variablen in der Stichprobe ausdrückt. Was lässt sich anhand des Resultats über den Zusammenhang in der Stichprobe sagen?

b) Kann Pierre auch für die gesamte Familie einen Zusammenhang annehmen? (Signifikanzniveau 5%)

c) Pierre hat außerdem festgestellt, dass das Alter der drei Personen positiv mit der TV-Nutzung ($r = 0{,}50$) und negativ mit dem Vertrauen korreliert ($r = -0{,}50$). Muss er befürchten, dass der in a) gefundene Zusammenhang (in der Stichprobe) nur eine Scheinkorrelation ist?