

M24 Statistik 1: Wintersemester 2024 / 2025

# Vorlesung 04: Korrelation

Prof. Matthias Guggenmos

Health and Medical University Potsdam





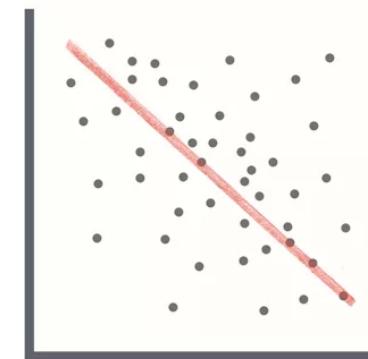
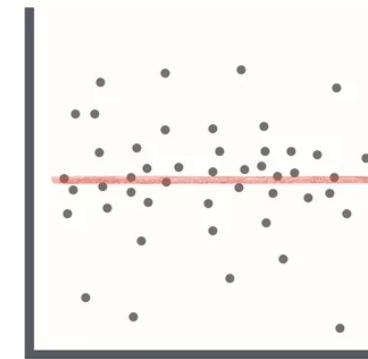
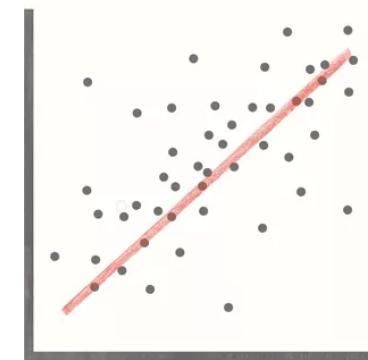
Die bisherige Auswertung der Beobachtungsstudie hat Evidenz sowohl für Hypothese 1 (mehr Zeit auf TikTok) als auch Hypothese 2 (erhöhte Entzündungswerte) erbracht. So einen richtigen Reim können Sie sich noch nicht auf das Ergebnis machen.

Ihre Neugier ist aber geweckt und Sie fragen sich: hängen vielleicht Ihre beiden abhängigen Variablen (TikTok-Zeit & Entzündungswerte) selbst miteinander zusammen? Steigen die Entzündungswerte mit zunehmender Online-Zeit auf der Plattform TikTok?

TikTok-Zeit  $\longleftrightarrow$  Entzündungswerte

# Was sind Zusammenhänge?

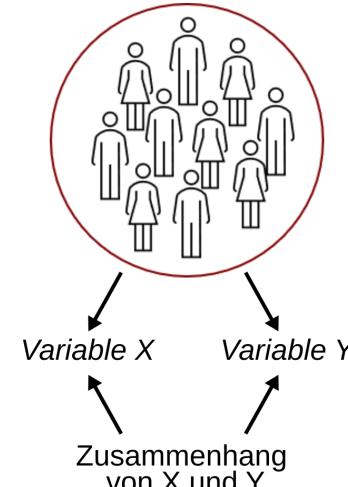
- Ein **Zusammenhang** beschreibt zu welchem Grad zwei Variablen (Merkmale) systematisch miteinander in Verbindung stehen
- Zusammenhänge bilden die Essenz der psychologischen Forschung –durch sie versuchen wir die Mechanik der menschlichen Psyche zu verstehen:
  - Fördert Ausdauersport das **psychische Wohlbefinden**?
  - Helfen **Psychotherapiestunden** bei der Überwindung einer **Depression**?
  - Wirkt sich **Bildschirmzeit** nachteilig auf die **Schlafqualität** aus?
  - Steigt durch **kindliche Frühförderung** die Wahrscheinlichkeit für einen **akademischen Bildungsabschluss**?
- Arten von Zusammenhängen:
  - Linearer Zusammenhang (Pearson-Korrelation)
  - Rangkorrelation
  - Lineare Regression (Unterscheidung von UV und AV)
  - Kontingenzkoeffizient



# Zusammenhänge versus Unterschiede

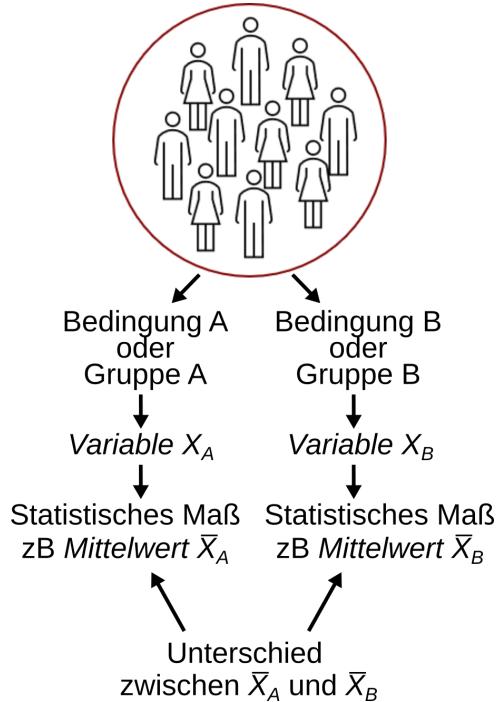
## ■ Zusammenhänge:

- Zugrunde liegen zwei Variablen  $X$  und  $Y$  in einer Gruppe.
- Verglichen werden Einzelwerte der Variablen  $X$  und  $Y$ .



## ■ Unterschiede:

- Zugrunde liegt eine Variable  $X$  in zwei Gruppen/Bedingungen  $\Rightarrow X_A, X_B$ .
- Verglichen werden statistische Maße, die für jede Gruppe/Bedingung auf Basis der Variable berechnet wurden (z.B. Mittelwerte  $\bar{x}_A$  vs.  $\bar{x}_B$ ).



## ■ Aber: häufig lassen sich **Unterschiedshypothesen** in **Zusammenhangshypothesen** überführen:

- Unterscheidet sich die akademische Leistung von Rauchern und Nichtrauchern?  
 $\Rightarrow$  Zusammenhang **Zahl der Zigaretten pro Tag** und **akademische Leistung**
- Unterscheidet sich Medienkonsum von Depressiven und Kontrollen?  
 $\Rightarrow$  Zusammenhang **Depressivität** und **Medienkonsum**
- Zusammenhangshypothesen sind häufig das “schärfere statistische Schwert”, da eine willkürliche und verlustbehaftete Einteilung einer Variablen in Kategorien vermieden wird.

# Kovarianz

# Kovarianz

- Ziel: mathematische Größe, die zum Ausdruck bringt, wie stark die Variation zweier Variablen miteinander in Zusammenhang steht.
- Wir haben bereits eine Größe für die Variation *einer Variable* – die Varianz:

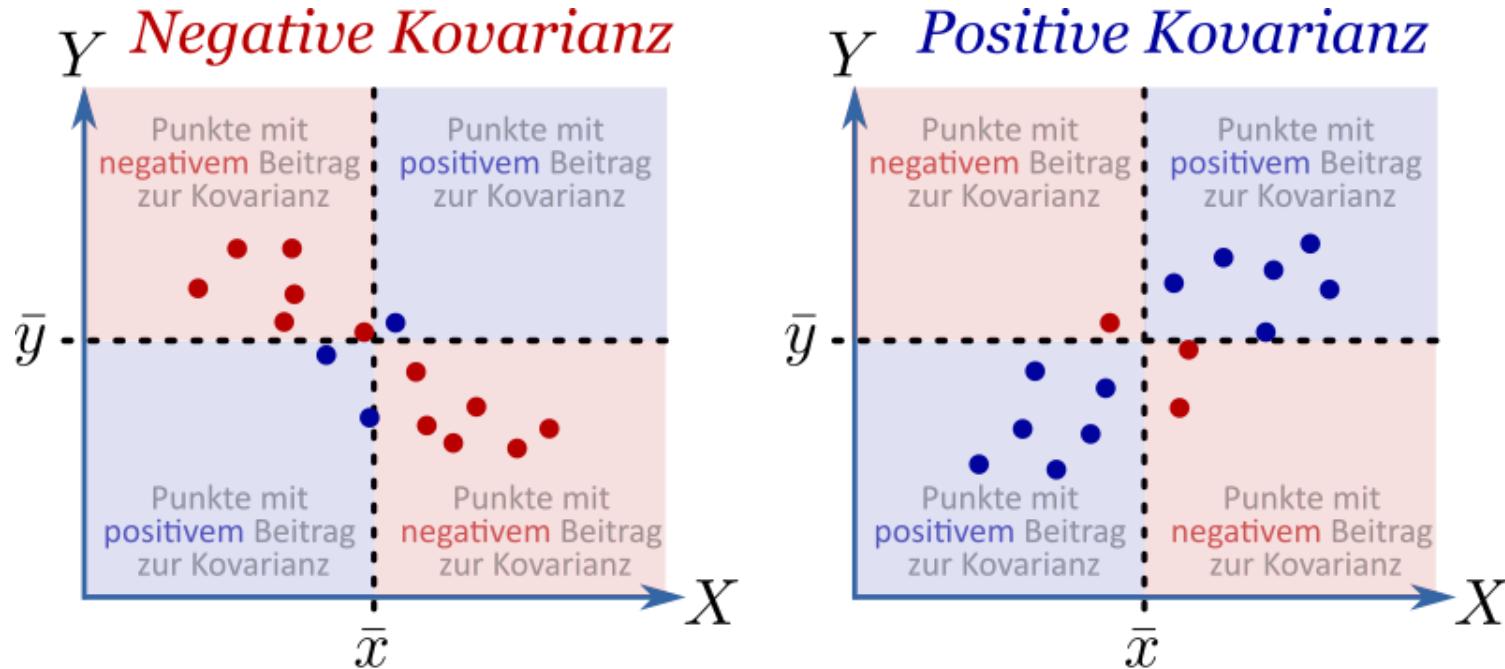
$$\hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})$$

- Die Varianz gibt an, wie stark eine Variable  $X$  um ihren Mittelwert  $\bar{x}$  schwankt.
- Analog berechnet die **Kovarianz**, wie stark die *gemeinsame Schwankung zweier Variablen um ihren jeweiligen Mittelwert* ist:

Kovarianz:  $\hat{Cov}(X, Y) = \hat{\sigma}_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

- Im Zentrum der Kovarianz steht die Erkenntnis, dass das **mathematische Produkt** zweier Abweichungsvariablen – hier die Abweichungen  $(x_i - \bar{x})$  und  $(y_i - \bar{y})$  vom Mittelwert – angibt, wie stark die beiden Abweichungen gleichsinnig variieren (ko-variieren).

# Kovarianz

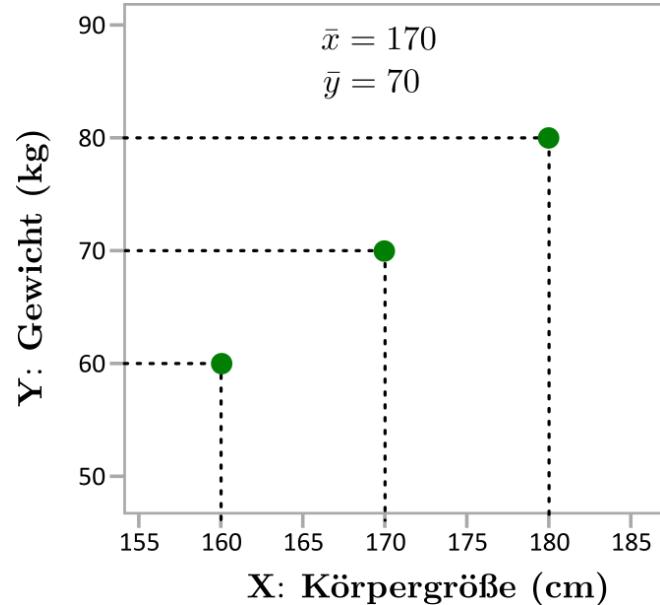


Intuition für positive Kovarianz:

- Sind zwei zusammengehörige Datenpunkte  $x_i$  und  $y_i$  **größer als der Mittelwert**, sind sowohl  $(x_i - \bar{x})$  als auch  $(y_i - \bar{y})$  **positiv**, und damit auch das Produkt  $(x_i - \bar{x})(y_i - \bar{y})$  **positiv**.
- Sind zwei zusammengehörige Datenpunkte  $x_i$  und  $y_i$  **geringer als der Mittelwert**, sind sowohl  $(x_i - \bar{x})$  als auch  $(y_i - \bar{y})$  **negativ**, und damit das Produkt  $(x_i - \bar{x})(y_i - \bar{y})$  wieder **positiv**.

Die Intuition für eine negative Kovarianz funktioniert ähnlich.

# Kovarianz: Größe-Gewicht-Beispiel



$$\begin{aligned}
 \hat{Cov}(X, Y) &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \\
 &= \frac{1}{2} [(160 - 170)(60 - 70) + (170 - 170)(70 - 70) + (180 - 170)(80 - 70)] = \\
 &= \frac{1}{2} [(-10) \cdot (-10) + 0 \cdot 0 + 10 \cdot 10] = \\
 &= \frac{1}{2} \cdot 200 = 100 \text{ [cm} \cdot \text{kg]}
 \end{aligned}$$

- Problem: die Kovarianz hängt von den Einheiten ab!
- Wird im Beispiel die Körpergröße  $X$  in der Einheit Meter angegeben, so lautet die Kovarianz:

$$\begin{aligned}
 \hat{Cov}(X, Y) &= \frac{1}{2} [(1.60 - 1.70)(60 - 70) + (1.70 - 1.70)(70 - 70) + (1.80 - 1.70)(80 - 70)] = \\
 &= \frac{1}{2} [(-0.10) \cdot (-10) + 0 \cdot 0 + 0.10 \cdot 10] = \frac{1}{2} \cdot 2 = 1 \text{ [m} \cdot \text{kg}]
 \end{aligned}$$

- Kovarianzen sind also **nicht vergleichbar, wenn sich Einheiten unterscheiden**, und erst recht nicht, wenn sich die Variablen unterscheiden.

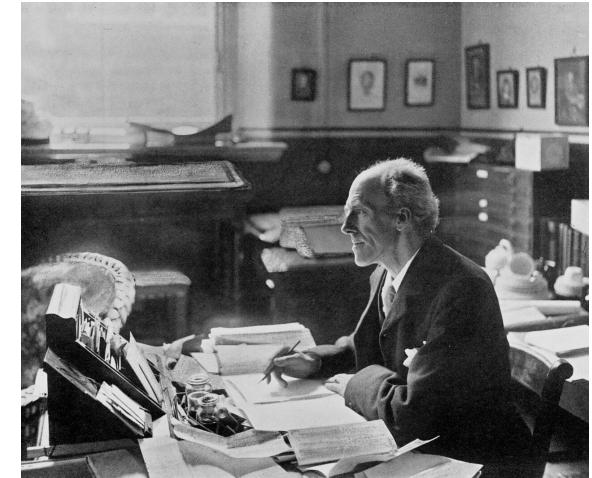
# Pearson-Korrelation

# Pearson-Korrelation

- Die **Pearson-Korrelation** schafft Abhilfe für das Problem der mangelnden Vergleichbarkeit.
- Der Schlüssel: die Kovarianz wird mit den **Standardabweichungen**  $\hat{\sigma}_X$  und  $\hat{\sigma}_Y$  beider Variablen normalisiert.
- Formel der Pearson-Korrelation:

$$\hat{\rho} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \dots = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}) \sum_{i=1}^n (y_i - \bar{y})}}$$

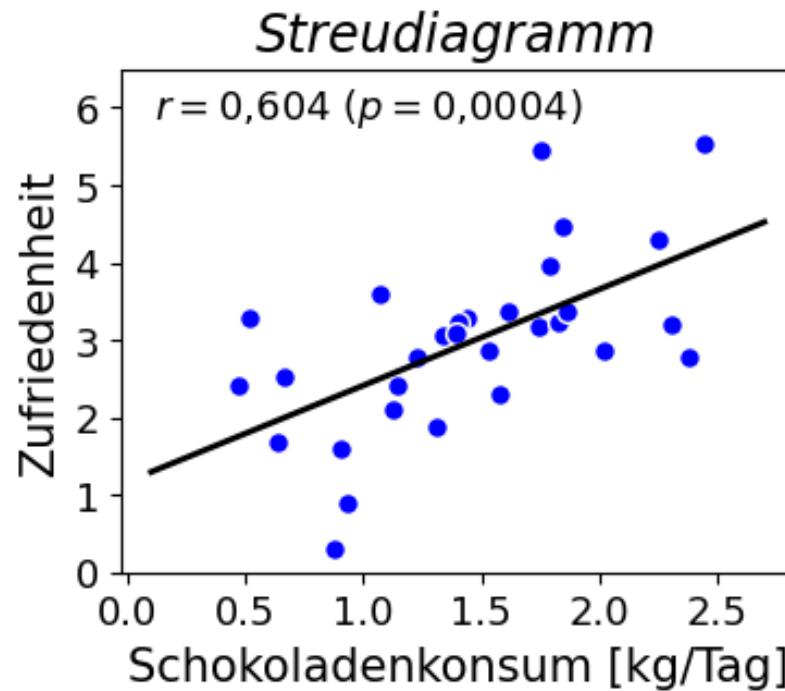
- Durch das Teilen durch die Standardabweichungen  $\sigma_X$  und  $\sigma_Y$  werden die Einheiten herausgekürzt – die Korrelation ist also eine **einheitslose Größe**.
- Die Korrelation kann Werte zwischen  $-1$  (perfekter negativer Zusammenhang) und  $+1$  (perfekter positiver Zusammenhang) annehmen.
- Hinweise:
  - Wir verwenden in der Korrelationsformel Größen ohne Besselkorrektur (d.h. ohne  $\wedge$ ).
  - In wissenschaftlichen Publikationen wird der Korrelationskoeffizient  $\hat{\rho}$  häufig vereinfacht mit  $r$  bezeichnet (z.B.  $r = 0.32$ ). Grund: die Formeln von inferentieller und nicht-inferentieller Korrelation sind identisch.



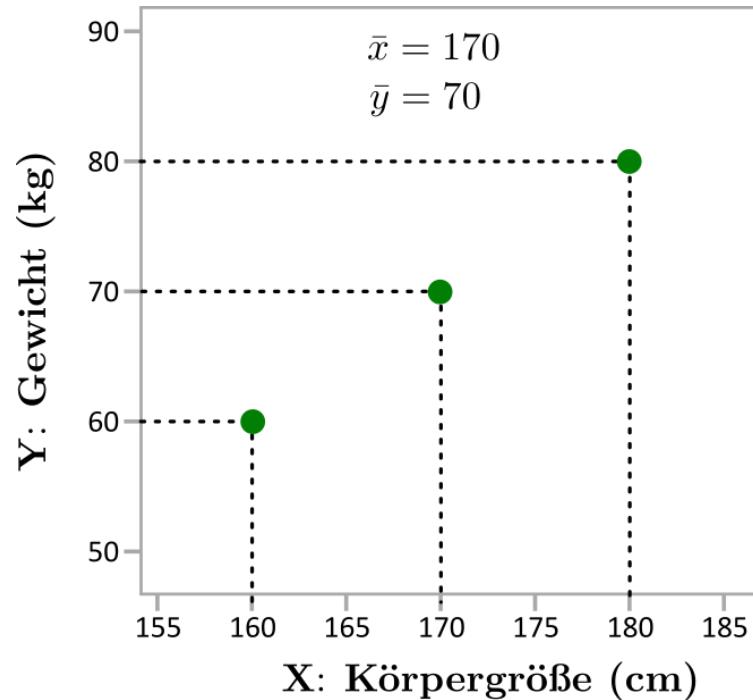
Der britische Mathematiker Karl Pearson in seinem Büro im Jahr 1910. Neben dem Korrelationskoeffizienten verdanken wir Pearson viele andere statistische Konzepte wie die Hauptkomponentenanalyse oder den p-Wert. Später wurden seine Ansichten zu Eugenik kritisch hinterfragt.  
Bildnachweis<sup>1</sup>

# Das Streudiagramm

- Zusammenhänge zweier Variablen werden häufig in Form eines **Streudiagramms** (engl. *scatter plot*) dargestellt.
- Bei der Korrelation ist es dabei willkürlich, welche Variable auf der x- und y-Achse liegt.
- Häufig wird zusätzlich zur “Punktwolke” auch eine **Regressionsgerade** dargestellt, sowie die Stärke des Zusammenhangs ( $r=..$ ,  $p=..$ ).



# Pearson-Korrelation: Größe-Gewicht-Beispiel



Kovarianz (ohne Besselkorrektur):

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \stackrel{\text{(prüfe nach)}}{=} \frac{200}{3} [\text{cm} \cdot \text{kg}]$$

Und berechnen nun die Korrelation  $\hat{\rho} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$

$$\sigma_X = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{3} [(-10)^2 + 0^2 + 10^2]} = \sqrt{\frac{200}{3}} [\text{cm}]$$

$$\sigma_Y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{\frac{1}{3} [(-10)^2 + 0^2 + 10^2]} = \sqrt{\frac{200}{3}} [\text{kg}]$$

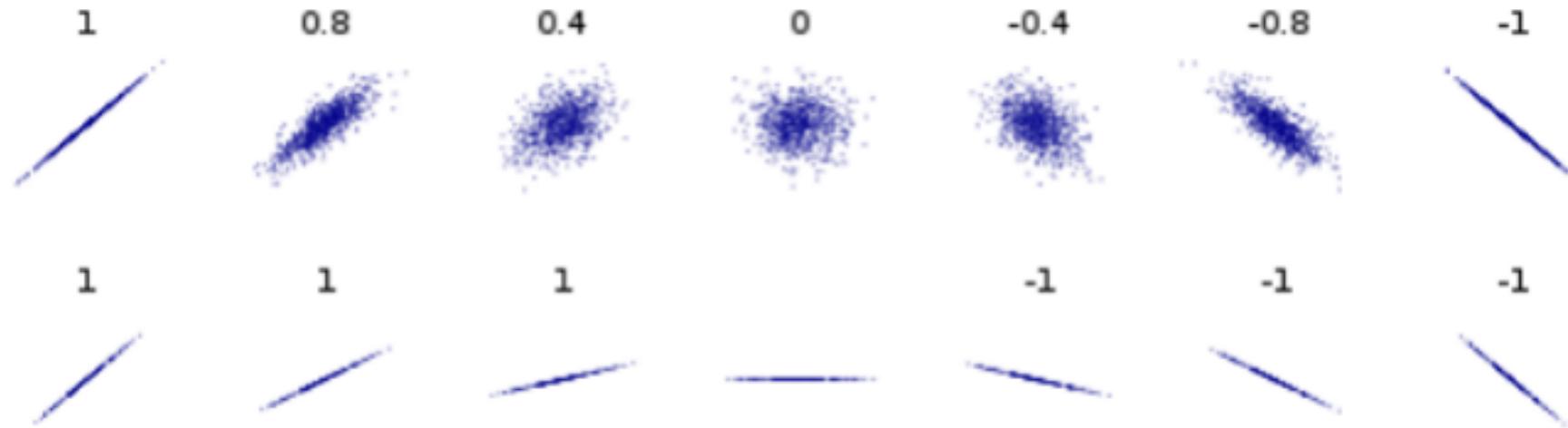
$$\hat{\rho} = \frac{\frac{200}{3} [\text{cm} \cdot \text{kg}]}{\sqrt{\frac{200}{3}} [\text{cm}] \cdot \sqrt{\frac{200}{3}} [\text{kg}]} = 1$$

- Die Einheiten kürzen sich beim Korrelationskoeffizienten  $\hat{\rho}$  also raus! Wird die Körpergröße statt in der Einheit Zentimeter etwa in Meter angeben, bleibt der Korrelationskoeffizient  $\hat{\rho}$  unverändert.

**Die Normalisierung mit der Standardabweichung sorgt dafür, dass die Korrelation unabhängig von der Einheit ist!**

# Interpretation der Pearson-Korrelation

- Die Pearson-Korrelation zeigt an, wie linear der Zusammenhang zweier Variablen ausgeprägt ist.
- Die Pearson-Korrelation ist dabei nicht von der Steigung einer gedachten Gerade abhängig.



Korrelationskoeffizienten für verschiedene hypothetische Streudiagramme

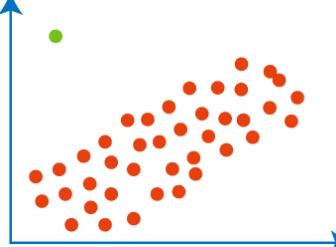
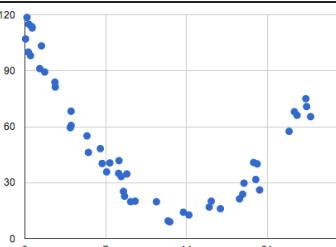
- Ein Zusammenhang zweier Variablen kann extrem schwach sein (z.B. so, dass eine Verdopplung von X nur einer 0.1%-Steigerung von Y entspricht) und dennoch kann die Korrelation stark sein (= nah an  $\pm 1$ ), wenn die Punkte exakt auf einer Geraden liegen.
- Auf einen Satz gemünzt kann man sagen:
  - Die Pearson-Korrelation misst, wie gut bivariate Daten durch eine Gerade abgebildet werden können.

<https://www.guessthecorrelation.com/>



# Voraussetzungen für das Berechnen der Pearson-Korrelation

- Die Pearson-Korrelation kann immer berechnet werden, solange beide Variablen aus Zahlenwerten bestehen.
- Es gibt jedoch weitere Kriterien, die für die Sinnhaftigkeit und Interpretierbarkeit der Pearson-Korrelation wichtig sind:

Kriterium	Falls Kriterium nicht erfüllt?	Beispiel	Mögliche Abhilfe?
Daten haben mindestens Intervallskalenniveau	<ul style="list-style-type: none"> <li>Keine Aussage über Linearität des untersuchten Zusammenhangs möglich</li> <li>Korrelation nicht interpretierbar</li> </ul>	<p>What Is Your Educational Background?</p> <ul style="list-style-type: none"> <li><input type="radio"/> 1 - Elementary</li> <li><input type="radio"/> 2 - High School</li> <li><input type="radio"/> 3 - Undegraduate</li> <li><input type="radio"/> 4 - Graduate</li> </ul>	Rangkorrelation
Keine Ausreißer	Korrelationskoeffizient kann massiv verzerrt sein		Ausreißer entfernen oder Rangkorrelation
Zusammenhang der Daten wird nicht durch nicht-linearen Anteil dominiert	<ul style="list-style-type: none"> <li>Pearson-Korrelation falsches Modell</li> <li>Linearität der Daten wird verzerrt wiedergegeben, da die Korrelation vom nicht-linearen Teil beeinflusst wird</li> </ul>		Komplexeres Modell, das den nicht-linearen Anteil berücksichtigt

# Mythen zur Pearson-Korrelation

- In vielen Quellen finden sich darüber hinaus **unzutreffende Behauptungen** zur Pearson-Korrelation:

	Behauptung	Fact
Mythos 1	Die Variablen müssen kontinuierlich sein	Pearson-Korrelation ist valide für diskrete Daten, solange diese mindestens Intervallskalenniveau aufweisen. Tatsächlich gibt es sogar eine Variante der Pearson-Korrelation, bei der beide Variablen binär sind (Phi-Koeffizient).
Mythos 2	Die Variablen müssen einen linearen Zusammenhang aufweisen	Gegenbeispiel: wenn Daten aus zufälligem Rauschen basieren, sind sie mit Sicherheit nicht linear verbunden – und dennoch gibt der Pearson-Koeffizient korrekterweise an, dass die Korrelation ungefähr 0 ist. Zusammenhänge in der Psychologie sind <i>sehr selten</i> eindeutig linear, dennoch kann es sinnvoll sein, die Pearson-Korrelation anzuwenden. Besser ist daher zu sagen (s. vorherige Folie), dass der Zusammenhang <b>nicht zu stark durch einen nicht-linearen Anteil dominiert</b> werden sollten.
Mythos 3	Die beiden Variablen müssen normalverteilt sein.	Der Pearson-Korrelationskoeffizient per se erfordert keine Normalverteilung der Variablen. Korrekt ist aber, dass die Daten für die <b>Berechnung eines p-Wertes auf Basis des t-Tests annähernd normalverteilt</b> (ganz korrekt: <i>bivariat normalverteilt</i> ) sein sollten. Wenn Normalverteilung nicht gegeben ist, können andere Signifikanztests (Permutation, Bootstrap) verwendet werden.
Mythos 4	Die Variablen müssen varianzhomogen sein	Varianzhomogenität (auch Homoskedastizität) meint, dass Y-Werte ähnliche Varianz in verschiedenen Abschnitten der X-Achse haben und umgekehrt. Hier gilt das gleiche wie bei Mythos 3: <b>Varianzhomogenität ist keine Voraussetzung für die Anwendung der Pearson-Korrelation per se, wohl aber für die Anwendung des t-Tests.</b>

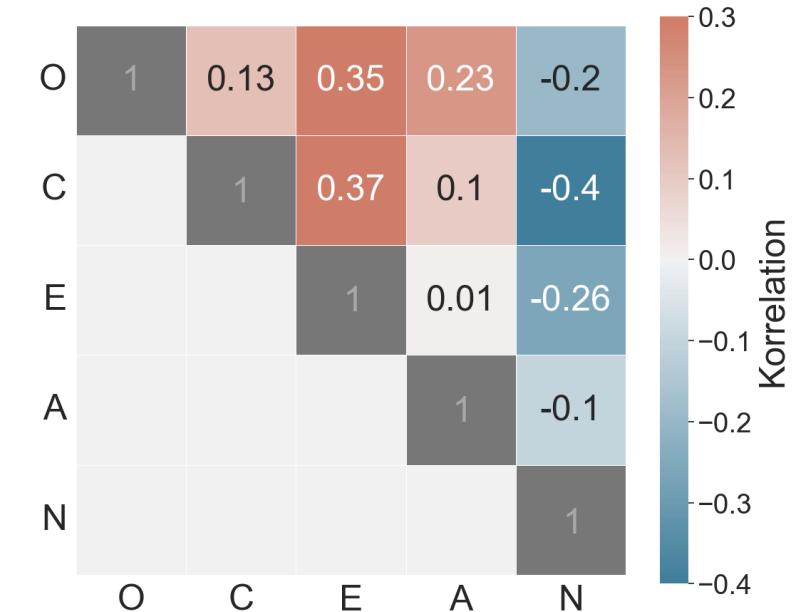
# Korrelationsmatrix

- Eine **Korrelation** bestimmt immer den Zusammenhang zwischen **zwei Variablen**.
- Gibt es mehr als zwei Variablen (z.B. die “Big Five”), bietet sich eine Darstellung aller paarweisen Korrelationen an – die **Korrelationsmatrix**.

Trait	O	C	E	A	N
Openness	1.00	.13	.35*	.23*	-.02
Conscientiousness		1.00	.37*	.10	-.40*
Extraversion			1.00	.01	-.26*
Agreeableness				1.00	-.10
Neuroticism					1.00

\* $p < .001$

Tabellarische Korrelationsmatrix. Sterne kennzeichnen häufig das Signifikanzniveau.



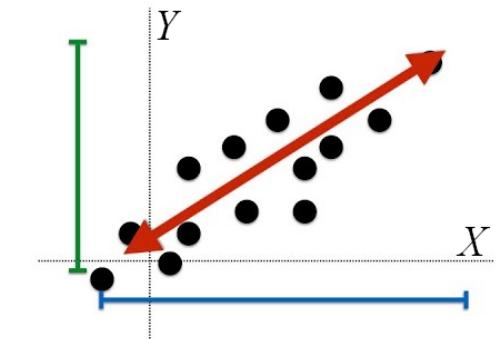
Korrelationsmatrix als “Heatmap” – die Einfärbung ist ein visuelles Hilfsmittel zur intuitiven und schnellen Erfassung der Korrelationsstruktur von mehreren Variablen.

- Die **Nebendiagonalelemente** sind der interessante Teil der Korrelationsmatrix, sie geben die Korrelationen verschiedener Variablen an.
- Die **Diagonalelemente**, also die Korrelationen von Variablen mit sich selbst, sind immer 1.

# Kovarianzmatrix

- Eine analoge Matrix-Darstellung gibt es auch für die **Kovarianz**.
- Im Unterschied zur Korrelationsmatrix sind die Diagonalelemente der Kovarianzmatrix nicht 1, sondern geben die **Varianz** der Variable an.
- Sei  $\mathbf{X}$  (beachte Fettschrift) ein Vektor von  $n$  Variablen  $X_1 \dots X_n$ , so ist die zugehörige Kovarianzmatrix  $\hat{Cov}(\mathbf{X})$ :

$$\hat{Cov}(\mathbf{X}) = \begin{pmatrix} \hat{Var}(X_1) & \hat{Cov}(X_1, X_2) & \cdots & \hat{Cov}(X_1, X_n) \\ \hat{Cov}(X_2, X_1) & \hat{Var}(X_2) & \cdots & \hat{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Cov}(X_n, X_1) & \hat{Cov}(X_n, X_2) & \cdots & \hat{Var}(X_n) \end{pmatrix}$$



$$\begin{pmatrix} \hat{Var}(X) & \hat{Cov}(X,Y) \\ \hat{Cov}(X,Y) & \hat{Var}(Y) \end{pmatrix}$$

- Im Gegensatz zur Korrelationsmatrix ist die Kovarianzmatrix selten das “Endprodukt” einer Analyse, sondern meist ein Zwischenschritt in fortgeschritteneren statistischen Analysen wie der **Hauptkomponentenanalyse**.

# Rangkorrelationen

# Was ist eine Rangkorrelation?

- Eine Alternative zur Pearson-Korrelation sind **Rangkorrelationsmaße**, die nicht die Linearität eines Zusammenhangs, sondern die **Monotonie des Zusammenhangs** bemessen.
- Rangkorrelationen werden in der Regel aus zwei Gründen angewendet:
  - **Theoretischer Grund:** Man ist tatsächlich an der Monotonie — und nicht der Linearität — eines Zusammenhangs interessiert. Dies ist häufig der Fall, wenn die Daten tatsächlich als ordinalskalierte Ränge – d.h. ohne interpretierbare Abstände – vorliegen.
  - **Praktischer Grund:** Eigentlich ist die Linearität das primäre Interesse, aber es sind entweder die Annahmen der Pearson-Korrelation verletzt (Daten nicht intervallskaliert) oder die Pearson-Korrelation ist aus anderen Gründen ungeeignet (Ausreißer). Die Rangkorrelation ist die “Notfallopption”.
- Für Rangkorrelationsmaße spielt lediglich die Reihenfolge (“Rang”) der Daten eine Rolle und nicht die spezifischen Werte.
  - Ähnlich wie beim Median ist es genau diese Eigenschaft, die Rangkorrelationsmaße unanfällig für Ausreißer macht.



Students	Maths	Science
A	35	24
B	20	35
C	49	39
D	44	48
E	30	45

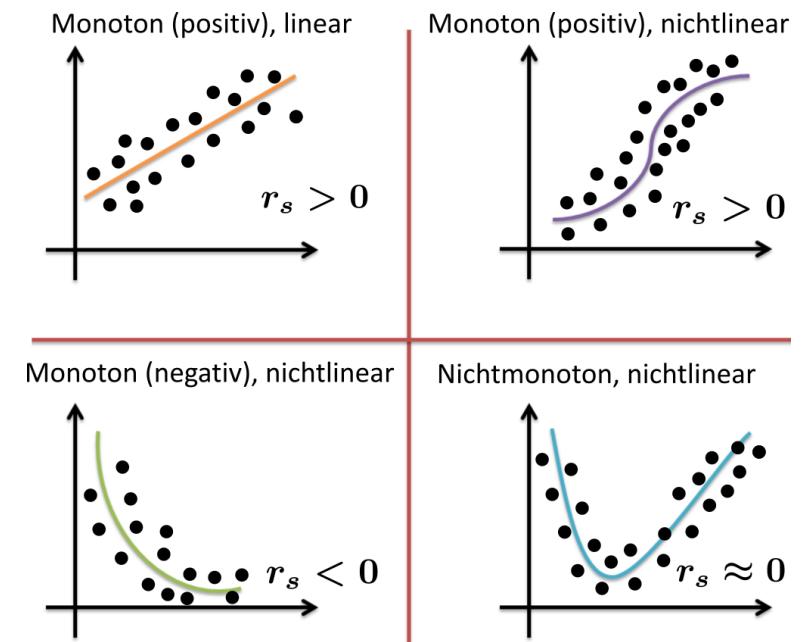
# Spearman-Korrelation

- Das mutmaßlich häufigste Rangkorrelationsmaß ist die **Spearman-Korrelation**  $\hat{\rho}_s$ .
- Die Spearman-Korrelation ist identisch zur Pearson-Korrelation, wenn die Variablen  $X$  und  $Y$  als Ränge  $R(X)$  und  $R(Y)$  vorliegen:

$$\hat{\rho}_s = \frac{Cov(R(X), R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}}$$

wobei  $\sigma_{R(X)}$  und  $\sigma_{R(Y)}$  die Standardabweichungen der Ränge von  $X$  und  $Y$  sind.

- Wie die Pearson-Korrelation nimmt die Spearman-Korrelation Werte zwischen  $-1$  und  $+1$  an:
  - Ein positiver Wert impliziert eine positiven monotonen Zusammenhang
  - Ein negativer Wert impliziert eine negativen monotonen Zusammenhang
  - Ein Wert nahe bei  $0$  impliziert einen schwachen (oder keinen) monotonen Zusammenhang



# Berechnung von Rängen

- Liegen die Variablen  $X$  oder  $Y$  nicht als Ränge vor, müssen sie zunächst in Ränge  $R(X)$  bzw.  $R(Y)$  umgewandelt werden:
  1. Werte der Variable sortieren
  2. (Unnormierte) Ränge zuordnen
  3. Normierung: Gleiche Werte erhalten den Mittelwert ihrer Ränge
  4. (Optional) Variablen in ihre ursprüngliche Reihenfolge bringen

Ausgangswerte		Werte sortieren und Ränge bilden				(Optional) Rücksortieren			
Index	Wert	Index	Wert	Rang	Normiert	Index	Wert	Rang	Normiert
1	1,5	5	1,0	1	1	1	1,5	2,5	
2	1,5	1	1,5	2	(2 + 3)/2	2	1,5	2,5	
3	4,0	2	1,5	3	= 2,5	3	4,0	5,0	
4	3,0	4	3,0	4	4	4	3,0	4,0	
5	1,0	3	4,0	5	5	5	1,0	1,0	
6	5,0	6	5,0	6	(6 + 7)/2	6	5,0	6,5	
7	5,0	7	5,0	7	= 6,5	7	5,0	6,5	
8	9,5	8	9,5	8	8	8	9,5	8,0	

- Die Tabelle gibt die Rangberechnung einer Variablen an (z.B. X) — für die andere Variable muss das analoge Prozedere durchgeführt werden.

# Kendalls Tau

- Eine Alternative Rangkorrelation zu Spearman ist **Kendalls Tau**  $\hat{\tau}$ .
- Kendalls Tau vergleicht inwieweit die Rangfolge *aller* Paare  $(x_i, x_j)$  mit der Rangfolge *aller* Paare  $(y_i, y_j)$  übereinstimmt.
- Dazu wird die Zahl der konkordanten (übereinstimmenden) und diskordanten (nicht übereinstimmenden) Paare gezählt.

$$\hat{\tau} = \frac{K - D}{K + D}$$

**K** ist die Anzahl der konkordanten Paare  
und **D** die Anzahl der diskordanten Paare.

## Beispiel

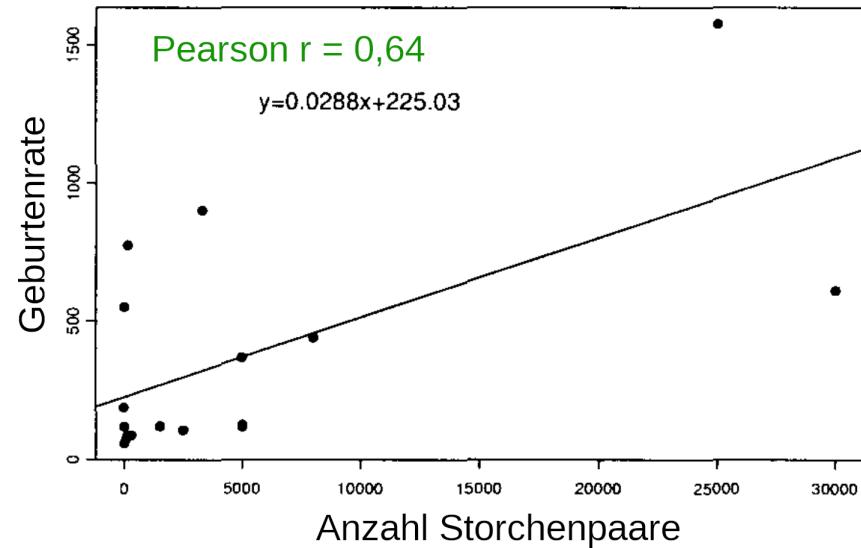
$$X = \begin{pmatrix} 9 \\ 3 \\ 7 \\ 5 \end{pmatrix} \quad Y = \begin{pmatrix} 18 \\ 7 \\ 8 \\ 21 \end{pmatrix}$$

- Die Paare von X sind:  $(9, 3), (9, 7), (9, 5), (3, 7), (3, 5), (7, 5)$
- Die Paare von Y sind:  $(18, 7), (18, 8), (18, 21), (7, 8), (7, 21), (8, 21)$
- Die Paare  $(x_1 = 9, x_3 = 7)$  und  $(y_1 = 18, y_3 = 8)$  wären **konkordant**, da die Rangfolge des X-Paars  $(x_1 > x_3)$  gleich der Rangfolge des entsprechenden Y-Paars  $(y_1 > y_3)$  ist.
- Die Paare  $(x_1 = 9, x_4 = 5)$  und  $(y_1 = 18, y_4 = 21)$  wären **diskonkordant**, da die Rangfolge des X-Paars  $(x_1 > x_4)$  ungleich der Rangfolge des entsprechenden Y-Paars  $(y_1 < y_4)$  ist.

Insgesamt gibt es im Beispiel 4 konkordante Paare und 2 diskordante Paare (prüfe nach!), daher gilt:

$$\hat{\tau} = \frac{K - D}{K + D} = \frac{4 - 2}{4 + 2} = \frac{2}{6} = 0.333..$$

# Ausreißer: Der Storch bringt die Babys zur Welt ( $p = 0.008$ )



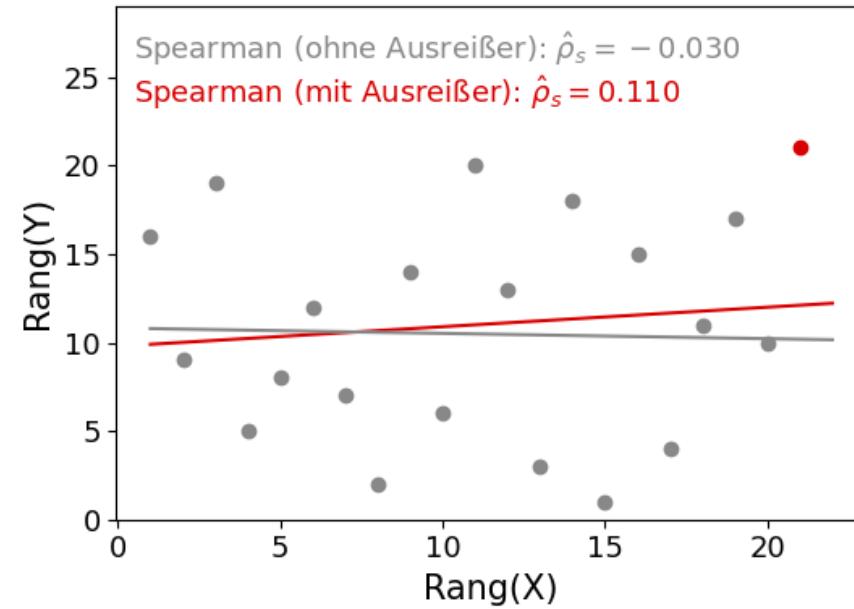
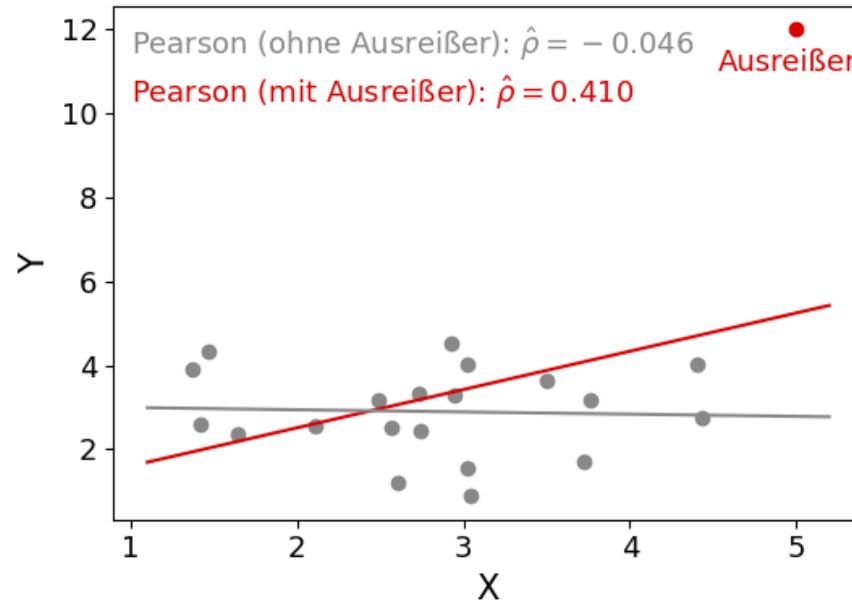
Land	Fläche (km <sup>2</sup> )	Störche (Paare)	Menschen (10 <sup>6</sup> )	Geburtenrate (10 <sup>3</sup> /Jahr)
Albanien	28 750	100	3.2	83
Belgien	30 520	1	9.9	87
Bulgarien	111 000	5000	9.0	117
Dänemark	43 100	9	5.1	59
Deutschland	357 000	3300	78	901
Frankreich	544 000	140	56	774
Griechenland	132 000	2500	10	106
Holland	41 900	4	15	188
Italien	301 280	5	57	551
Österreich	83 860	300	7.6	87
Polen	312 680	30 000	38	610
Portugal	92 390	1500	10	120
Rumänien	237 500	5000	23	23
Spanien	504 750	8000	39	439
Schweiz	41 290	150	6.7	82
Türkei	779 450	25 000	56	1 576
Ungarn	93 000	5000	11	124

[http://www3.math.uni-paderborn.de/~agbiehler/sis/sisonline/struktur/jahrgang21-2001/heft2/Langfassungen/2001-2\\_Matth.pdf](http://www3.math.uni-paderborn.de/~agbiehler/sis/sisonline/struktur/jahrgang21-2001/heft2/Langfassungen/2001-2_Matth.pdf)

- Im gezeigten Beispiel wird der Wert der **Pearson-Korrelation** fast ausschließlich durch zwei Ausreißer (Türkei und Polen) dominiert. Der Zusammenhang wird dadurch überschätzt.
- Ein Rangkorrelationsmaß wie die **Spearman-Korrelation** wäre hier eine sinnvolle Alternative – obwohl eigentlich die Linearität des Zusammenhangs von primärem Interesse ist.

# Rangkorrelationen: Robust gegen Ausreißer

- Beispiel Pearson vs. Spearman:



- Ein einziger Ausreißer verändert die Pearson-Korrelation im Beispiel von  $\hat{\rho} = -0.046$  nach  $\hat{\rho} = 0.410$ .
- Demgegenüber ist die Spearman-Korrelation “robuster” gegenüber dem Ausreißer – sie verändert sich “lediglich” von  $\hat{\rho}_s = -0.030$  nach  $\hat{\rho}_s = 0.110$ .
- Intuition: während der Wert des Ausreißers deutlich über dem zweithöchsten Y-Wert liegt, ist der Rang nur um 1 höher.

# Der Phi-Koeffizient

- Spezialfall: beide Variablen haben nur zwei Ausprägungen (sind also *dichotom*).
- Darstellbar in der **Vierfeldertafel**:

	male	female	$\Sigma$
nonsmoker	a	b	$a+b$
smoker	c	d	$c+d$
$\Sigma$	$a+c$	$b+d$	n

- In die vier Felder werden die absoluten Häufigkeiten (idR Anzahl Versuchspersonen) der jeweiligen Variablen-Kombination eingetragen.
- Optional: In der letzten Zeile/Spalte die Summe.
- $a + b + c + d$  muss sich zur Stichprobengröße  $n$  addieren.
- Frage: hängen Raucherstatus und Geschlecht zusammen? Die Antwort liefert der **Phi-Koeffizient**:

Phi-Koeffizient: 
$$\hat{\phi} = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

- Wichtig: das Vorzeichen hängt davon ab, in welcher Reihenfolge die beiden Variablen in die Vierfeldertafel eingetragen werden.
  - Übung für zu Hause: das Vorzeichen von  $\phi$  dreht sich um, wenn die Spalten male/female vertauscht werden, und ebenso, wenn die Zeilen smoker/nonsmoker vertauscht werden. Warum?
- Wie alle Korrelationskoeffizienten hat auch der Phi-Koeffizient einen Wertebereich von  $[-1; 1]$ .

# Der Phi-Koeffizient

- Der Phi-Koeffizient ist identisch mit der Pearson-Korrelation und der Spearman-Korrelation, wenn jeweils beide dichotomen Variablen  $X$  und  $Y$  mit den Werten 0/1 kodiert werden.

$$\hat{\phi} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad \text{mit} \quad X = \{0, 1\}, Y = \{0, 1\}$$

- Hier ist für das Vorzeichen entscheidend, welche Ausprägung als 0 und welche als 1 definiert wird (analog der Eintrage-Reihenfolge in die Vierfeldertafel).
- In der Praxis findet der Phi-Koeffizient wenig Anwendung, da der Zusammenhang zumeist intuitiver in Form von Häufigkeiten berichtet wird (z.B. Raucherhäufigkeit bei Männern versus Frauen).
- Zwei Vorteile hat der Phi-Koeffizient jedoch:
  - Es muss keine Festlegung erfolgen, ob die Raucherhäufigkeit zwischen den Geschlechtern, oder die Geschlechterhäufigkeit zwischen Rauchern und Nichtrauchern berichtet wird.
  - Der Phi-Koeffizient kann in Metaanalysen mit anderen Studien verglichen werden, die eine ähnliche Fragestellung mit einer linearen Pearson-Korrelation bemessen haben.
- Zudem macht der Phi-Koeffizient in konzeptioneller Hinsicht den Punkt zu Beginn der Vorlesung deutlich, dass jeder Unterschied (z.B. Raucherhäufigkeit bei Männern versus Frauen) auch als Zusammenhang (Zusammenhang von Geschlecht und Raucherstatus) formuliert werden kann.

# Phi-Koeffizient: Beispiel

		Haben Sie ein Haustier?		Gesamt
		ja	nein	
Geschlecht	männlich	2	8	10
	weiblich	12	11	23
Gesamt		14	19	33

$$\hat{\phi} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{2 \cdot 11 - 8 \cdot 12}{\sqrt{(2+8)(12+11)(2+12)(8+11)}} = \\ = \frac{22 - 96}{\sqrt{10 \cdot 23 \cdot 14 \cdot 19}} = \frac{-74}{\sqrt{61180}} = -0.299$$

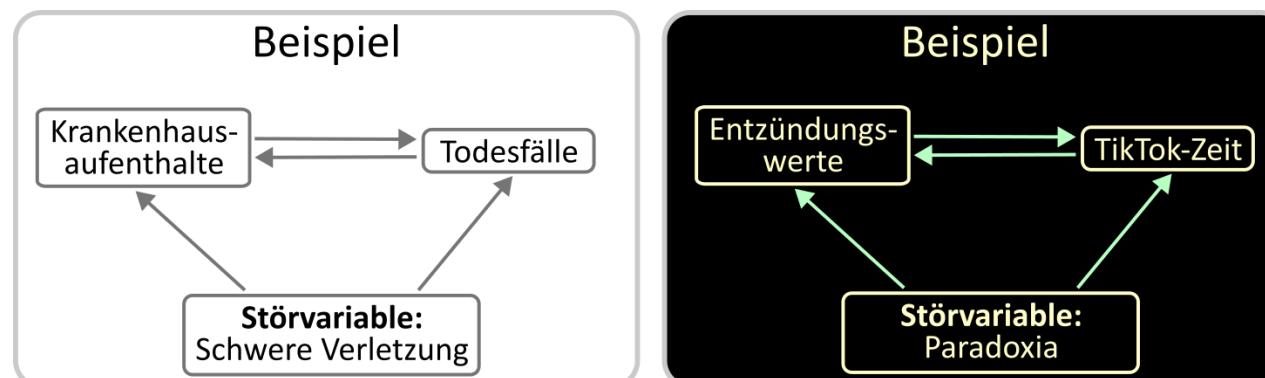
# Interpretation von Korrelationen



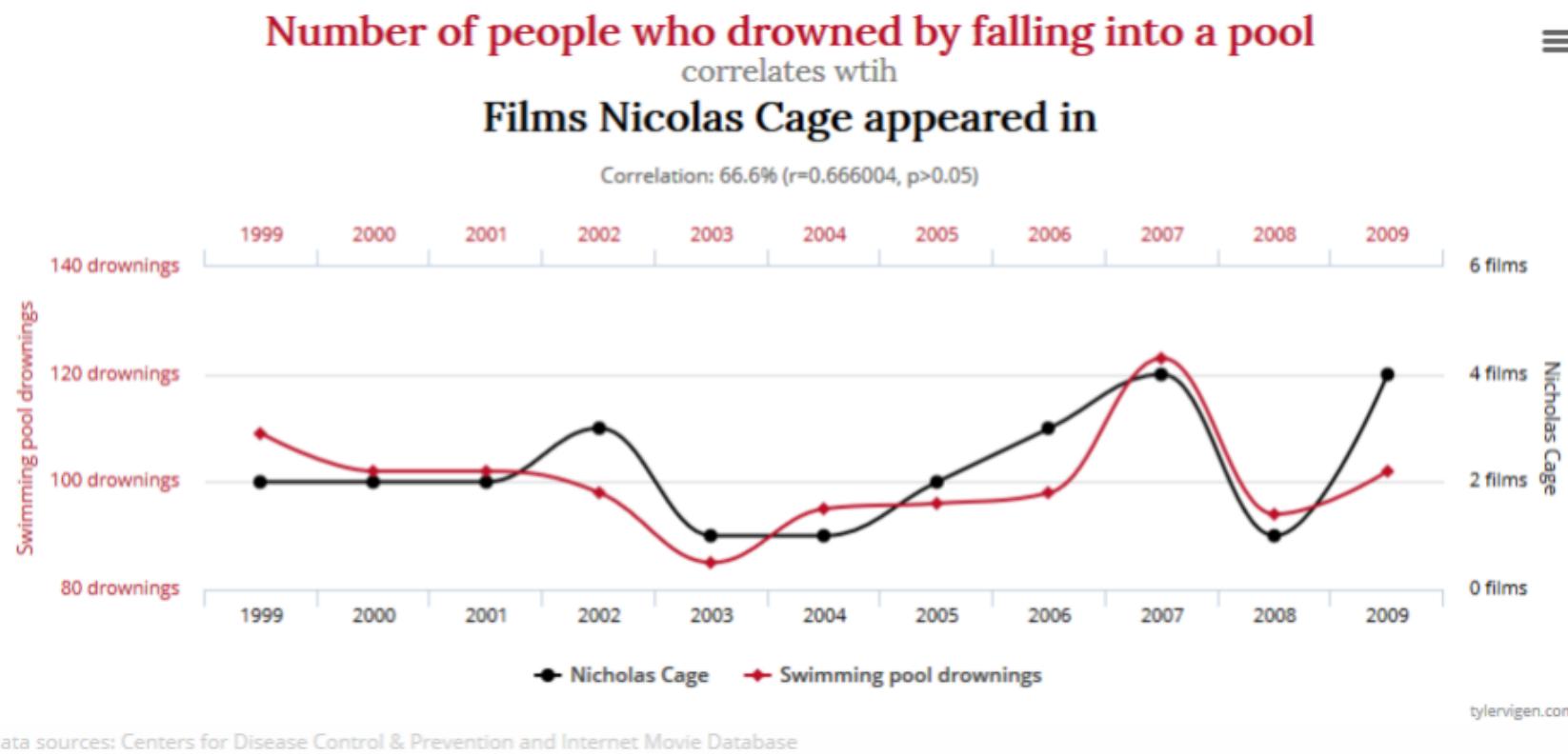
Bildnachweis<sup>2</sup>

# Korrelation und Kausalität

- Eine positive oder negative Korrelation zeigt an, dass zwei Variablen zusammenhängen – und auch, dass man *eine Variable aus der anderen (zu einem gewissen Grad) vorhersagen kann*.
- Dies bedeutet jedoch nicht, dass sich die Variablen auch kausal bedingen
- Einerseits können Korrelationen zufällig zustande kommen – die Wahrscheinlichkeit von solchen irrtümlichen Befunden untersuchen wir noch genauer beim Thema Signifikanztestung.
- .. andererseits können Korrelationen durch dritte Variablen (**Störvariablen**) verursacht sein.
- Umgekehrt können Störvariablen auch tatsächlich vorhandene Zusammenhänge unterdrücken – man spricht dann von **Suppression**; der Korrelationskoeffizient unterschätzt in diesem Fall die tatsächliche Stärke des Zusammenhangs.
- Trotz aller Fallstricke bei der Interpretation gilt: eine Korrelation KANN ein **Indiz für Kausalität** sein.
  - Wo Kausalität da auch Korrelation, sofern diese nicht durch Störvariablen unterdrückt ist.



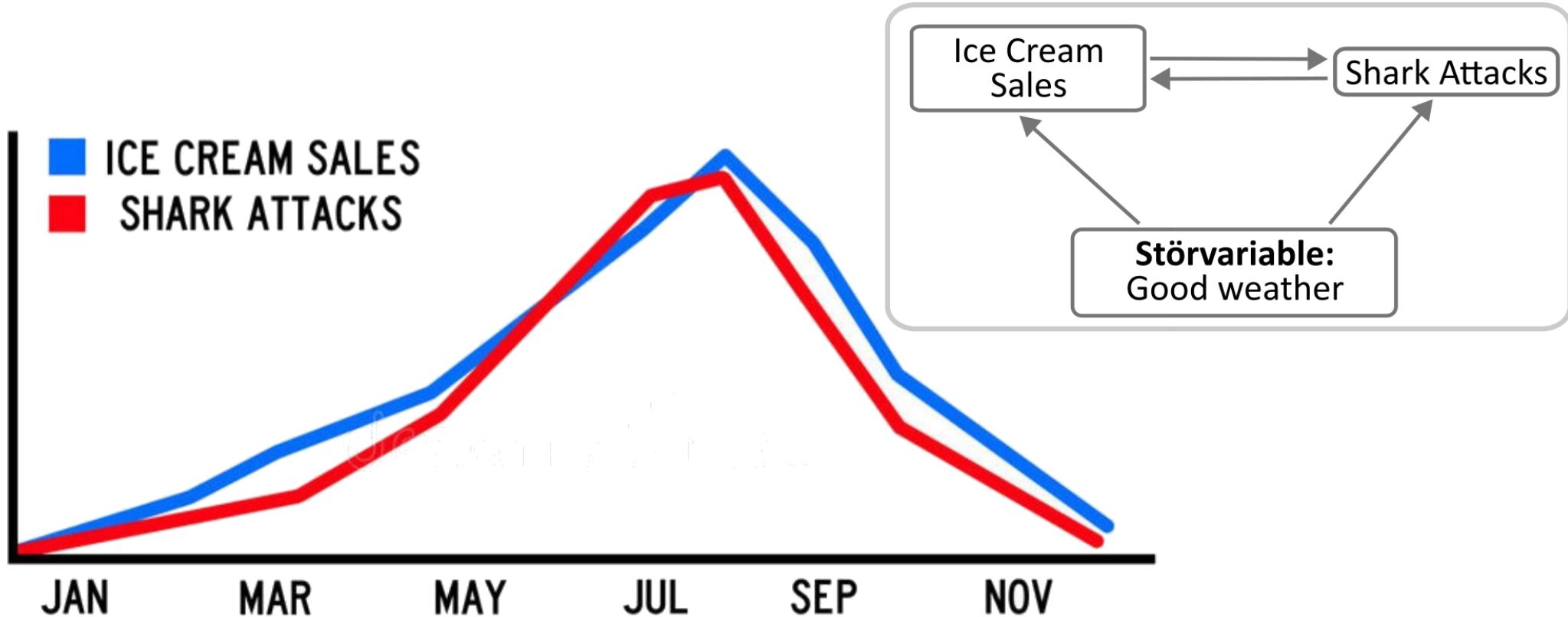
# Korrelation und Kausalität



tylervigen.com

Spurious correlations

# Korrelation und Kausalität



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

# Wann ist ein Korrelationskoeffizient groß oder klein?

- Pauschal schwer zu beantworten
- In der psychologischen Literatur hat sich folgende Konvention/Nomenklatur nach Jacob Cohen<sup>3</sup> eingebürgert:

Korrelation (r)	Konvention
0.1 bis 0.3	“kleiner” Effekt
0.3 bis 0.5	“mittlerer” Effekt
> 0.5	“großer” Effekt

- Allerdings fügt Cohen im gleichen Artikel hinzu:

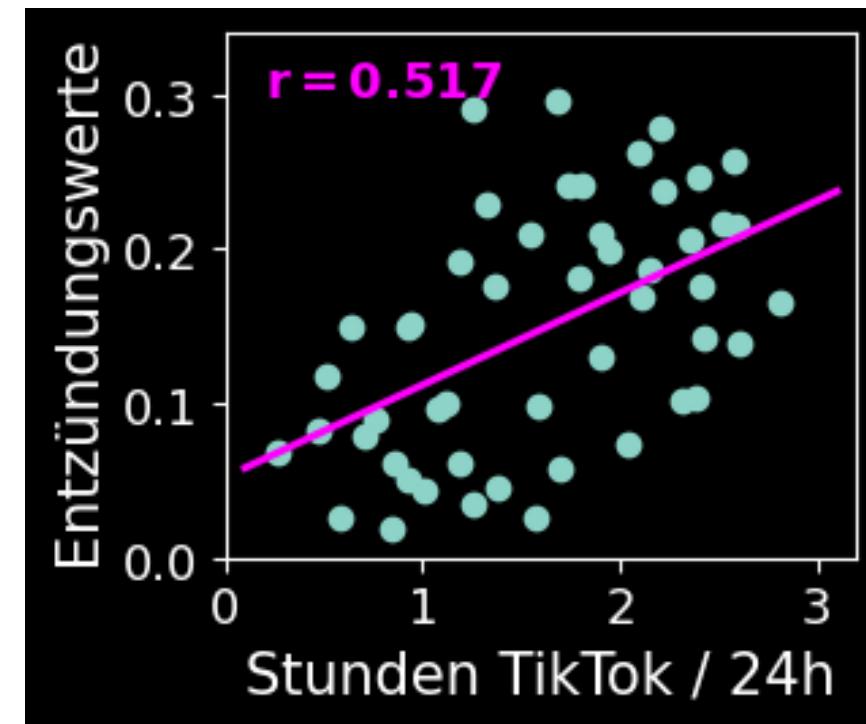
These proposed conventions were set forth throughout with much diffidence [Zurückhaltung], qualifications [Bedingungen], and invitations not to employ them if possible.

- Bei der Beurteilung sollte der Korrelationskoeffizient, wie jede Effektgröße, immer in Relation zu typischen Werten im jeweiligen Forschungsfeld gesetzt werden.

# [ Zusammenfassung ]

- Zusammenhänge werden mit **Kovarianz** und **Korrelation** untersucht.
- Im Gegensatz zur Kovarianz sind **Korrelationen unabhängig von den gewählten Einheiten**.
- Die **Pearson-Korrelation** misst die Linearität eines Zusammenhangs und gilt für intervallskalierte Daten.
- Für ordinalskalierte Variablen eignen sich **Rangkorrelationen** wie **Spearmans Rho** oder **Kendalls Tau**: sie messen die Monotonie eines Zusammenhangs.
- Sind beide Variablen binär, greift der **Phi-Koeffizient**.
- Correlation does not imply causation, Correlation does not imply causation, Correlation does not imply causation, Correlation does not...

Zurück zum Zusammenhang zwischen TikTok-Online-Zeit und Entzündungswerten. Beides sind kontinuierliche intervallskalierte Variablen, Sie können also den Pearson-Korrelationskoeffizienten anwenden. Sie stellen den Zusammenhang in Form eines **Streudiagramms** dar:



Faszinierend. Es gibt tatsächlich einen Zusammenhang beider Variablen. Kann das Ergebnis so interpretiert werden, dass TikTok sich auf physiologische Entzündungswerte auswirkt?

# Warum ist die Korrelation auf $-1$ bis $1$ beschränkt?

Darstellung der Korrelation nur mit (Ko)Varianzen:

$$\hat{\rho}_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{Cov(X, Y)}{\sqrt{Var(X)} \sqrt{Var(Y)}}$$

Die Korrelation sollte maximal ( $\hat{\rho} = 1$ ) sein, wenn  $X$  mit sich selbst korreliert wird ( $Y = X$ ). Zu berücksichtigen ist, dass die Kovarianz von  $X$  mit sich selbst gleich der Varianz ist:

$$\hat{\rho}_{XX} = \frac{Cov(X, X)}{\sqrt{Var(X)} \sqrt{Var(X)}} = \frac{Var(X)}{\sqrt{Var(X)} \sqrt{Var(X)}} = \frac{Var(X)}{Var(X)} = 1$$

Umgekehrt sollte die Korrelation maximal negativ sein ( $\hat{\rho} = -1$ ), wenn  $Y$  genau das Inverse von  $X$  ist, also  $Y = -X$ . Unter Berücksichtigung von  $Var(X) = Var(-X)$  gilt:

$$\hat{\rho}_{X(-X)} = \frac{Cov(X, -X)}{\sqrt{Var(X)} \sqrt{Var(-X)}} = \frac{-Cov(X, X)}{\sqrt{Var(X)} \sqrt{Var(X)}} = \frac{-Var(X)}{Var(X)} = -1$$



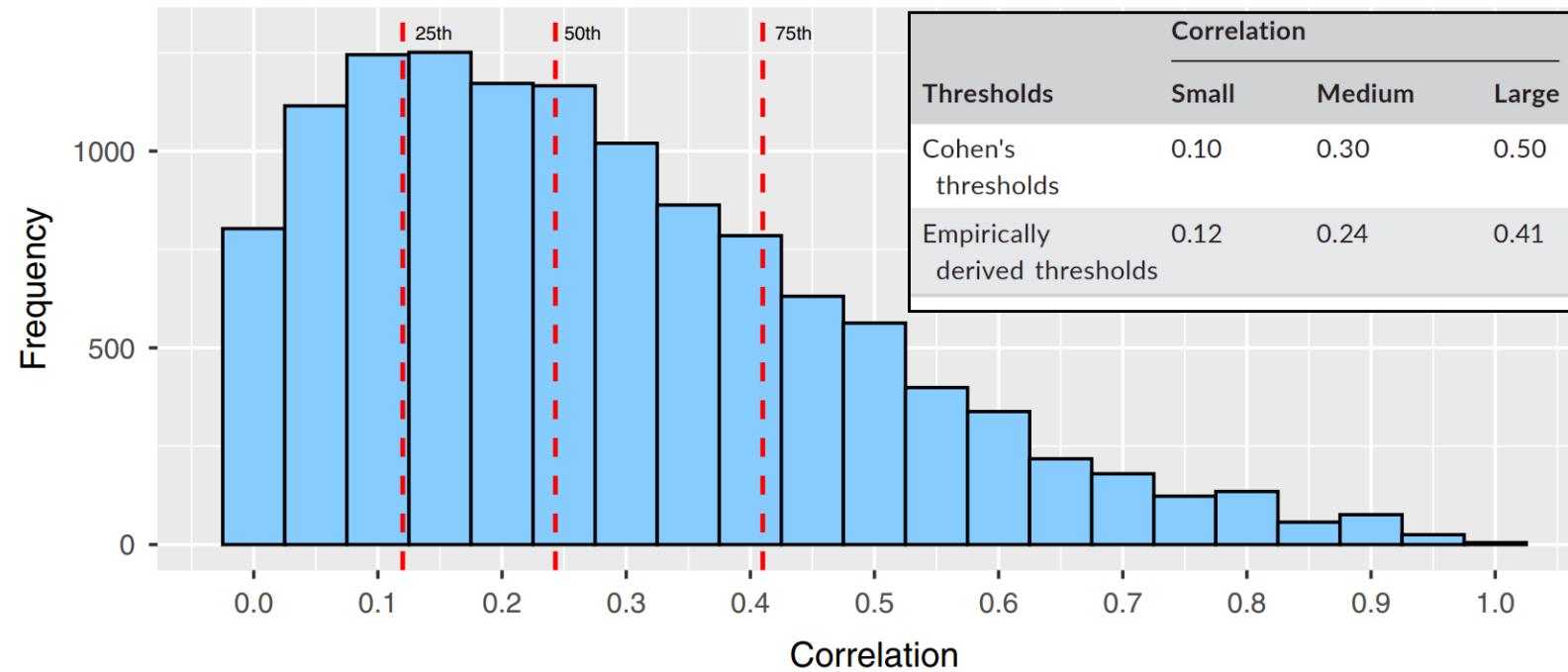
# Wann Spearman und wann Kendall?

- Beide Rangkorrelationskoeffizienten bestimmen die Monotonie eines Zusammenhangs.
- **Kendall ist robuster bei kleinen Stichproben** und ist in diesen Fällen bevorzugt.
- **Spearman ist etwas weniger sensitiv gegenüber Rangbindungen** (also wenn zwei Werte den gleichen Rang haben) und ist daher bevorzugt, wenn es viele Rangbindungen gibt<sup>7</sup>.
  - Beachte: auch bei den Kendall'schen Paaren gibt es Rangbindungen, und zwar dann, wenn die verglichenen Paare zwischen  $X$  und  $Y$  genau identisch sind.
  - Diese Paare sind weder konkordant noch diskordant und es gibt verschiedene Algorithmen diese Fälle zu berücksichtigen (hier nicht behandelt).
- **In Abwesenheit von Rangbindungen liefert Kendall präzisere Schätzungen** und ist in diesem Fall zu bevorzugen.<sup>8</sup>.
- In der Statistik wird Kendall häufig als “Default”-Rangkorrelation empfohlen<sup>9</sup>:
  - idR präzisere Schätzung des Populationsparameters
  - Standardfehler ist bekannt (für Spearman gibt es lediglich Approximationen<sup>10</sup>)
- In der Praxis ist aber Spearman der weitaus verbreiteter Korrelationskoeffizient — womöglich weil er idR größer als der Kendall-Koeffizient ist 😊.



# Wann ist ein Korrelationskoeffizient groß oder klein?

- In einer kürzlichen Metaanalyse von Lovakov Agadullina (2021)<sup>11</sup> wurden typische Effektstärken in der Sozialpsychologie untersucht:



- Erkenntnis: “mittlere” und “starke” Korrelationen (definiert als das 50%- und 75%- Quantil) sind in der Realität kleiner als von Cohen angenommen



# Fußnoten

1. <http://www.learn-stat.com/life-of-karl-peerson/>
  2. <https://www.patheos.com/blogs/tippling/2017/10/31/spooky-punkins-and-statistical-correlation/>
  3. Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences. New York, NY: Routledge Academic
  4. <https://real-statistics.com/correlation/biserial-correlation/>
  5. [https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Point-Biserial\\_and\\_Biserial\\_Correlations.pdf](https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Point-Biserial_and_Biserial_Correlations.pdf)
  6. Jacobs P, Viechtbauer W (2017) Estimation of the biserial correlation and its sampling variance for use in meta-analysis: Biserial Correlation. *Res Syn Meth* 8:161–180.
  - 7.
- Puth M-T, Neuhäuser M, Ruxton GD (2015) Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour* 102:77–84.
- Puth M-T, Neuhäuser M, Ruxton GD (2015) Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour* 102:77–84.
9. DC Howell (2012). Statistical Methods for Psychology. Wadsworth.
  10. <https://stats.stackexchange.com/questions/18887/how-to-calculate-a-confidence-interval-for-spearmans-rank-correlation>
  11. Lovakov A, Agadullina ER (2021) Empirically derived guidelines for effect size interpretation in social psychology. *Eur J Soc Psychol* 51:485–504.