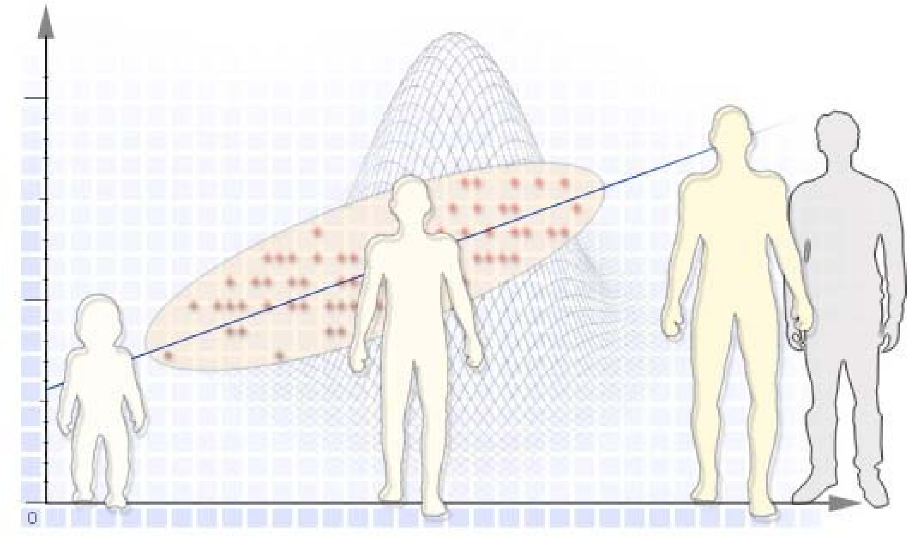
KOR - Korrelation



Lernziele und Überblick

In dieser Lerneinheit werden Sie Maßzahlen für den linearen Zusammenhang zweier Merkmale kennenlernen.

Lernziele

Nach dem Durcharbeiten dieser Lerneinheit sollten Sie in der Lage sein

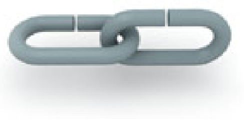
- den Korrelationskoeffizienten zu berechnen
- Korrelationen zu interpretieren
- den Zusammenhang zweier Merkmale angemessen mit der linearen Korrelation zu charakterisieren
- die Korrelation aus einem Streudiagramm abzuschätzen
- die Gefahr der Fehlinterpretation von Scheinkorrelationen zu erläutern.

Gliederung der Lerneinheit

Zeitbedarf und Umfang

Für die Durcharbeitung dieser Lerneinheit benötigen Sie ca. 150 Minuten und für die Übungen mit der Statistiksoftware **R** ca. 60 Minuten.

1 Einleitung



Jetzt wird es wissenschaftlich! Das Studium von Zusammenhängen zwischen Variablen – ihrer Korrelation – hat eine wesentliche Rolle bei der Begründung der modernen Vererbungslehre gespielt. Charles Darwin (1809–1882) benutzt den Begriff, um allgemeine Prinzipien für Wachstum und Selektion zu erklären.

> *Was geht das uns an?*

Wir werden sehen, wie sich lineare Beziehungen zwischen Variablen leicht durch eine einzige Zahl angeben lassen. Zur Berechnung benötigen wir Mittelstufen-Mathematik. Also, keine Bange und viel Spaß beim Lernen.

In der folgenden Abbildung sehen Sie noch einmal eine Punktwolke, ähnlich wie wir sie in Lerneinheit *„ZHA - Zusammenhänge“* auch gesehen haben. Beachten Sie die Boxplots, die wie bei Kontingenztafeln die jeweiligen Randverteilungen darstellen. Bei beiden Boxplots stimmt der eingetragene Median fast mit der Koordinatenachse überein, es gibt also jeweils etwa gleich viele positive und negative Werte, gemeinsam nehmen die Variablen aber fast nur Werte im I. und im III. Quadranten ein.

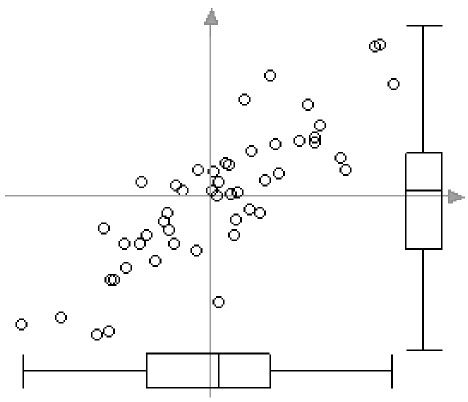
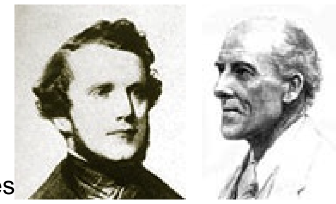


Abb.:
Punktwolke und Boxplots

Die Korrelationsanalyse untersucht Zusammenhängen zwischen zwei gleichwertigen kardinalen Merkmalen. Am Anfang der Analyse sollte stets ein Streudiagramm der Daten angelegt werden. Aus Lage und Form der dargestellten Punktwolke lassen sich die Stärke und die Richtung des Zusammenhangs der Merkmale ablesen. Das Streudiagramm liefert erste Hinweise über eine mögliche Abhängigkeit zwischen Merkmalen. Dieser bildliche Eindruck soll im Folgenden durch geeignete Maßzahlen geprüft und präzisiert werden.

2 Der Korrelationskoeffizient von BRAVAIS-PEARSON



Wir haben in der Lerneinheit *„VAR - Varianz und Standardabweichung"* gesehen, dass Angaben zur Lage alleine einen Datensatz nicht ausreichend charakterisieren. Die Variabilität der Daten liefert weitere wichtige Informationen.

Beim Beschreiben des Zusammenhangs zweier Merkmale verhält es sich ebenso. Die Mittelwerte von X und Y alleine geben keine ausreichende Information. Wir müssen uns noch zusätzliche Information über die Variabilität der Wertepaare (X, Y) verschaffen.

Geeignete Maßzahlen werden im Folgenden zunächst wieder an Beispielen erläutert. Sie werden die Kovarianz und den Korrelationskoeffizienten kennenlernen.

Der Korrelationskoeffizient $r_X Y$ von Auguste Bravais und Karl Pearson ist eine Maßzahl der Stärke und der Richtung (Art) des linearen Zusammenhanges zweier Merkmale.

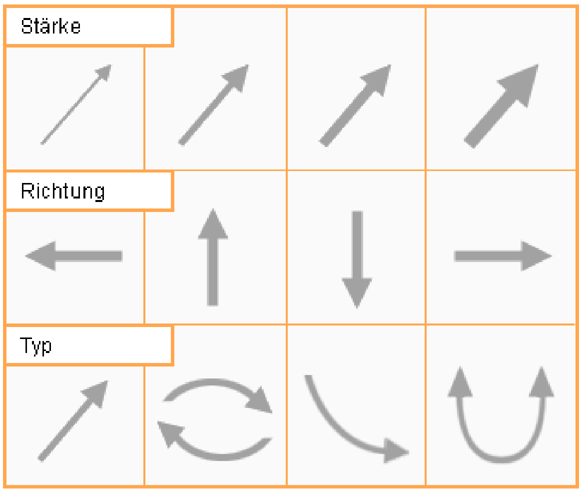


Abb.: Stärke, Richtung Typ einer Korrelation

Die Linearität sollte im Streudiagramm überprüft werden. Wir zeigen einige Beispiele linearer Zusammenhänge verschiedener Stärke und Richtung. Zunächst werden Sie am Beispiel wichtige Bausteine zur Konstruktion unserer Maßzahl kennenlernen. Ähnlich wie bei der Berechnung der empirischen Varianz werden Abweichungen der Einzelwerte vom Mittelwert betrachtet. Schauen Sie sich die Zahlenwerte in Ruhe an.

2.1 Beispiel Werbung und Umsatz im Weinhandel

Beginnen wir mit einem Beispiel, dessen Umfeld im Rahmen des Studienmoduls schon häufiger aufgetaucht ist – die Weinhandlung Maestro.

Beispiel

Umsatz und Werbeausgaben

Für die zehn Abteilungen des Ihnen bereits bekannten Weinfachgeschäftes Maestro sind im Oktober 2022 Daten über den Umsatz einer bestimmen Sorte Prosecco sowie über die Werbeausgaben für diese Sorte in der nachfolgenden Tabelle zusammengestellt:

Abteilungen	Werbung X	Umsatz Y
1	20.000	90.000
2	28.000	120.000
3	13.000	50.000
4	15.000	90.000
5	12.000	60.000
6	25.000	120.000
7	30.000	140.000
8	10.000	50.000
9	8.000	30.000
10	17.000	110.000

Tab.: Abteilungen, Werbung und Umsatz

Es soll statistisch untersucht werden, ob zwischen Werbungskosten X und Umsatz Y von n = 10 Abteilungen ein statistischer Zusammenhang besteht, wie stark er ausgeprägt ist und welche Richtung er besitzt.

Für die Zusammenhangsanalyse ist es wichtig, die folgenden Überlegungen anzustellen:

- Die statistische Einheit ist eine Abteilung;
- Die Erhebungsmerkmale sind die zwei kardinalen Merkmale:
  - Werbungskosten X,
  - Umsätze Y.

Die mittleren Werbungskosten und Umsätze sind:

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{178000}{10} = 17800\,€$$

$$\overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{860000}{10} = 86000\,€$$

In der folgenden Tabelle werden die Abweichungen der Einzelwerte vom Mittelwert und einige weiteren Hilfsgrößen bestimmt. Abweichungsquadrate kennen wir schon von der Berechnung der Varianz, neu sind die Produkte der X- und Y-Abweichungen.

Diese Produkte haben natürlich etwas mit der gemeinsamen Variabilität von X und Y zu tun. Haben wir es mit einem positiven Zusammenhang zu tun, dann besitzen die meisten der X- und Y-Abweichungen das gleiche Vorzeichen (positive Produkte), bei einem negativen Zusammenhang ergeben sich überwiegend negative Produkte und wenn die Vorzeichen der Produkte in etwa gleichmäßig verteilt sind, deutet wenig auf einen linearen Zusammenhang hin.

Was finden Sie in der folgenden Tabelle vor?

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	20.000	90.000	2.200	4.000	8.800.000	4.840.000	16.000.000
2	28.000	120.000	10.200	34.000	346.800.000	104.040.000	1.156.000.000
3	13.000	50.000	-4.800	-36.000	172.800.000	23.040.000	1.296.000.000
4	15.000	90.000	-2.800	4.000	-11.200.000	7.840.000	16.000.000
5	12.000	60.000	-5.800	-26.000	150.800.000	33.640.000	676.000.000
6	25.000	120.000	7.200	34.000	244.800.000	51.840.000	1.156.000.000
7	30.000	140.000	12.200	54.000	658.800.000	148.840.000	2.916.000.000
8	10.000	50.000	-7.800	-36.000	280.800.000	60.840.000	1.296.000.000
9	8.000	30.000	-9.800	-56.000	548.800.000	96.040.000	3.136.000.000
10	17.000	110.000	-800	24.000	-19.200.000	640.000	576.000.000
Σ	178.000	860.000	0	0	2.382.000.000	531.600.000	12.240.000.000

Tab.: Abweichungen der Einzelwerte vom Mittelwert und weitere Hilfsgrößen

Betrachten Sie bitte noch einmal die Rechenschritte der Tabelle, wir werden diese gleich weiter verwenden.

2.2 Streudiagramm (Interpretation) zum vorhergehenden Beispiel

Zur Berechnung von Maßzahlen für den Zusammenhang zwischen Werbung und Umsatz werden wir die Ergebnisse der Hilfsrechung des vorherigen Abschnitts verwenden.
Zunächst betrachten wir aber unbedingt die grafische Darstellung der Werte, wir wollen doch wissen, wovon wir reden!

Tragen wir in das Streudiagramm unserer Daten die Mittelwertlinien von X und Y ein, können wir leicht erkennen, wie sich beide Variablen bezüglich ihrer Mittelwerte verhalten. In unserem Beispiel ist es so, dass überdurchschnittliche Umsätze meist mit überdurchschnittlichen Werbungskosten einhergehen, für Werte unterhalb des Durchschnitts gilt das Entsprechende. Nur zwei Wertepaare bilden Ausnahmen von der Regel, dabei sind die Y-Abweichungen aber geringfügig.

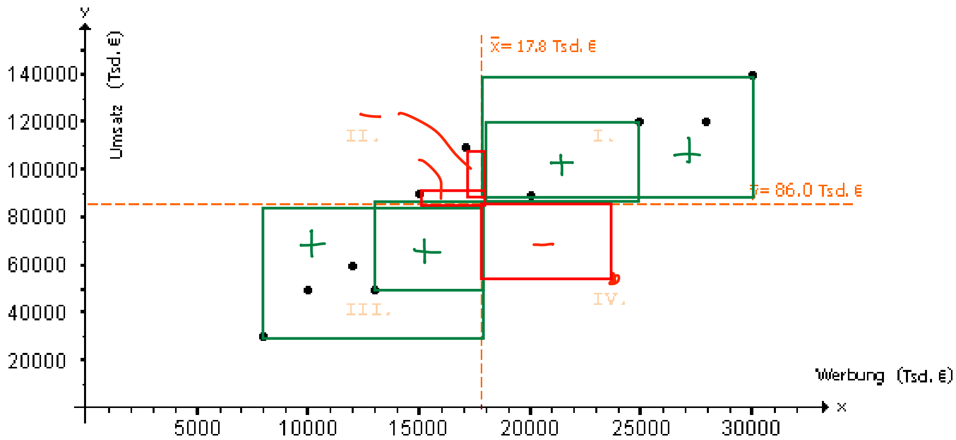


Abb.: Streudiagramm des Umsatzes und Werbungsausgaben mit jeweiligen Mittelwertlinien

Beachtenswert ist, dass gleichläufige Abweichungen stets positive Abweichungsprodukte, gegenläufige Abweichungen stets negative Abweichungsprodukte erzeugen. Diese Beobachtungen werden sowohl in der vorangehenden Tabelle als auch im Streudiagramm ersichtlich.

Aus dem gestreckten und steigenden Verlauf der Punktewolke ist zu erkennen, dass für die n = 10 Abteilungen zwischen den Werbungskosten X und dem Umsatz Y ein gleichläufiger linearer statistischer Zusammenhang besteht.

Jetzt verwenden wir die schon diskutierten Produkte der X- und Y-Abweichungen zur Quantifizierung. Wir mitteln diese Werte einfach und erhalten die Kovarianz.
(Über die Verwendung von n - 1 haben wir schon ausführlich in Lerneinheit VAR gesprochen.)

Kovarianz

Kovarianz

Die empirische Kovarianz s_{XY} der beobachteten Werbungskosten x_i und des zugehörigen Umsatzes y_i ist der aus den Daten $(x_i, y_i), i = 1, 2, \ldots, n$ berechnete Parameter

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{2382000000}{9}$$

$$= 264666666{,}667€^2$$

Die Kovarianz weist auf einen positiven statistischen Zusammenhang zwischen X und Y hin.

2.3 Erläuterungen zum BRAVAIS-PEARSON-Korrelationskoeffizienten

Im vorhergehenden Abschnitt haben wir den Begriff der Kovarianz eingeführt. Sie haben gesehen, dass positive Werte der Kovarianz auf einen positiven Zusammenhang hinweisen.

Bisher wissen wir aber nicht, wie die berechneten Werte zu beurteilen sind.

Exkurs: Die Kovarianz bei perfektem linearen Zusammenhang.

Auf jeden Fall hängt der Wertebereich der Kovarianz von der Variabilität der X und Y-Werte ab. Vielleicht können wir eine Standardisierung erreichen, wenn wir das mittlere Produkt der Abweichungen durch die zugehörigen Standardabweichungen s_X und s_Y dividieren. Zunächst betrachten wir die Standardabweichungen.

Hängen die Merkmale X und Y perfekt linear zusammen, dann gilt die Beziehung $Y = a + bX$. Wir erinnern uns an Lerneinheit *„VAR“*, Abschnitt 2.4. Da haben wir festgestellt, dass zwischen der Varianz von X und Y die folgende Beziehung besteht:

$s_{Y^2} = b^2 s_{X^2}$, bzw. $s_Y = |b| s_X$

Wir wollen untersuchen, was diese lineare Beziehung für die Kovarianz s_{XY} bedeutet.
Für die Kovarianz ergibt eine entsprechende Rechnung

$s_{XY} = b s_{x^2}$

Das folgt sofort, wenn Sie die Werte für Y durch $a + bX$ ersetzen.
Es gilt nämlich: $(x - \overline{x})(y - \overline{y}) = (x - \overline{x})(a + bx - [a + b\overline{x}]) = b(x - \overline{x})^2$

Jetzt berechen wir die standardisierte Kovarianz s_{XY} / (s_X x s_Y) und erhalten:

$$\frac{s_{XY}}{(s_X \cdot s_X)} = \frac{b\ s_X^2}{(s_X \cdot |b| s_X)} = \text{Vorzeichen}(b)$$

Also, bei einem perfekten linearen Zusammenhang hat die standardisierte Kovarianz entweder den Wert +1 oder –1, je nach dem Vorzeichen des Steigungskoeffizienten b. In der Wirklichkeit gibt es fast nie perfekte Beziehungen. Für alle Beziehungen zwischen perfektem positivem und negativem Zusammenhang liegt auch der Wert der standardisierten Kovarianz zwischen +1 und –1. Das kann man auch beweisen.

Korrelationskoeffizient = standardisierte Kovarianz

Jetzt wissen wir also, wie wir mit der Kovarianz umgehen sollten. Durch Standardisieren, man nennt das auch Normieren, wird sie in eine Größe mit Werten zwischen -1 und +1 überführt. Die standardisierte Kovarianz heißt Korrelationskoeffizient.

Berechnung des Korrelationskoeffizienten

Wenn wir die empirische Kovarianz mit dem Produkt der empirischen Standardabweichungen

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{531600000}{9}} = 7685{,}484\,€$$

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{1224000000}{9}} = 36878{,}178\,€$$

der Werbungskosten x_i und des Umsatzes y_i normieren, dann erhalten wir den dimensionslosen Korrelationskoeffizienten nach Bravais und Pearson von

$$r_{XY} = \frac{264666666{,}667\,€^2}{36878{,}178\,€ \cdot 7685{,}484\,€} \approx 0{,}93$$

Der berechnete Korrelationskoeffizient ist positiv und liegt nahe Eins.
Aus diesem Grund kann man ihn wie folgt interpretieren:

Zwischen den Werbungskosten X und den Umsätzen Y der betrachteten n = 10 Abteilungen eines Weinfachgeschäftes besteht ein ausgeprägter positiver linearer statistischer Zusammenhang. Demnach geht für die betrachteten Abteilungen in der Regel eine überdurchschnittliche Werbungsausgabe mit einem überdurchschnittlichen Umsatz bzw. eine unterdurchschnittliche Werbungsausgabe mit einem unterdurchschnittlichen Umsatz einher. Anscheinend hilft Werbung hier oder umsatzstarke Abteilungen können sich vielleicht mehr Werbung leisten.

2.4 Korrelationsanalyse (Beispiel)

Im nächsten Beispiel betrachten wir einen negativen Zusammenhang.

Beispiel

Beispiel für Korrelationsanalyse

Nur noch wenige können ein Gedicht auswendig aufsagen.

In der folgenden Tabelle sind die Ergebnisse eines Tests zur Gedächtnisleistung festgehalten. Es haben insgesamt 9 Studentinnen teilgenommen.

Der Test mit einer Studentin fand unter erschwerten Bedingungen statt: in einem mit sieben Personen gefüllten Büro, wobei sich vier unterhielten und eine telefonierte.

Anzahl auswendig zu lernender Wörter	5	10	15	20	25	30	35	40	45
Prozentsatz der davon behaltenen Wörter (%)	80	70	33	20	20	27	17	23	13

Tab.: Anzahl von auswendig gelernten und behaltenen Wörter

Es soll untersucht werden, ob zwischen den Angaben bezüglich der Anzahl aus einer Liste auswendig zu lernenden Wörtern X und dem Prozentsatz der davon behaltenen Wörter Y ein statistischer Zusammenhang besteht, wie stark er ausgeprägt ist und welche Richtung er besitzt.

Was vermuten wir eigentlich? Es könnte gut sein, dass das Erinnerungsvermögen begrenzt ist, so dass mit zunehmender Wortzahl der gemerkte Anteil abnimmt. Die vorliegenden Daten legen diese Vermutung nahe.

Lösung

Lösung

Die Mittelwerte der Variablen sind

$$\overline{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{225}{9} = 25 \text{ Wörter}$$

$$\overline{y} = \frac{1}{9}\sum_{i=1}^{9} y_i = \frac{303}{9} = 33{,}7\,\%$$

Wir berechnen wie im vorigen Beispiel die folgenden Zwischenergebnisse:

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	5	80	-20	46,33	- 926,67	400	2146,78
2	10	70	-15	36,33	- 545,00	225	1320,11
3	15	33	-10	- 0,67	6,67	100	0,44
4	20	20	-5	- 13,67	68,33	25	186,78
5	25	20	0	- 13,67	0	0	186,78
6	30	27	5	- 6,67	- 33,33	25	44,44
7	35	17	10	- 16,67	- 166,67	100	277,78
8	40	23	15	- 10,67	- 160,00	225	113,78
9	45	13	20	- 20,67	- 413,33	400	427,11
Σ	225	303	0	0	- 2170,00	1500	4704,00

Tab.: Zwischenergebnisse des Beispiels

Interpretation

Streudiagramm (Interpretation)

Aus der Punktewolke wird ersichtlich, dass die überdurchschnittliche Anzahl aus einer Liste auswendig zu lernender Wörter in der Regel mit dem unterdurchschnittlichen Prozentsatz der davon behaltenen Wörter und umgekehrt einhergehen.

Anders ausgedrückt: Je mehr Wörter auswendig zu lernen sind, desto geringer ist der Anteil der behaltenen Worte.

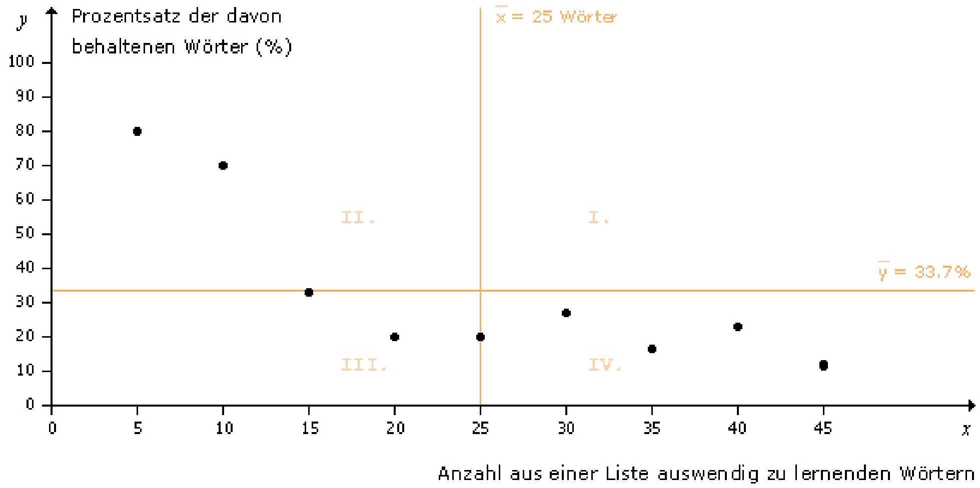


Abb.: Streudiagramm der Anzahl aus einer Liste auswendig zu lernender Wörter und des Prozentsatzes der davon behaltenen Wörter

Interpretation

Korrelationskoeffizient (Interpretation)

Wir erhalten einen Korrelationskoeffizienten nach A. Bravais und K. Pearson von

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{-271{,}25 \text{ Wort } \cdot \%}{13{,}69 \text{ Wort } \cdot \ 24{,}25\%} \approx -0{,}82$$

wobei empirische Kovarianz s_{XY}:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{-2170}{8} = -271{,}25 \text{ Wort} \cdot \%$$

empirische Standardabweichungen:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1500}{8}} = 13{,}69 \text{ Worte}$$

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{4704}{8}} = 24{,}25\%$$

Der berechnete Korrelationskoeffizient $r_{XY} = -0{,}82$ bestätigt den oben aus dem Streudiagramm gezogenen Schluss, dass der Korrelationskoeffizient eine ausgeprägte negative lineare Beziehung zwischen der Anzahl aus einer Liste auswendig zu lernender Wörter und dem Prozentsatz der davon behaltenen Wörter zum Ausdruck bringt.

2.5 Übung zur Korrelationsanalyse

Es gibt auch Situationen, da erwarten wir überhaupt keinen Zusammenhang. Eine solche finden Sie in der nächsten Übung.

Berechnen

Übung KOR-01

Studierende und Entfernung zur Hochschule

In der folgenden Tabelle werden Studierende, deren erreichte Punkte in einer Statistikklausur und die Entfernung ihres Wohnortes zur Hochschule aufgelistet.

	Person	Erreichte Punkte (X)	Entfernung in km (Y)
1	Mandy	88	5
2	Anna	94	13
3	Roland	85	6
4	Swenja	86	8
5	Alexander	82	12
6	Tanja	97	6
7	Irene	92	7
8	Edin	87	11
9	Karoline	84	12
10	Nikolas	90	15

Untersuchen Sie bitte, ob zwischen der Klausurnote der Studierenden X und der Entfernung des Wohnortes zur Hochschule Y ein statistischer Zusammenhang besteht, wie stark er gegebenenfalls ausgeprägt ist und welche Richtung er besitzt.

Lösung (Siehe Anhang)

Bearbeitungszeit: 15 Minuten

2.6 Kovarianz

Wie wir gesehen haben, eignet sich das Mittel der Produkte der Abweichungen zweier Merkmale von ihren jeweiligen arithmetischen Mitteln zur Quantifizierung der gemeinsamen Streuung. Diese wird Kovarianz genannt. Wie immer kommt auch eine Bestimmung mit Hilfe von Gewichten in Frage.

Definition

Kovarianz

Die empirische Kovarianz ist als beschreibende Kennzahl einer zweidimensionalen Verteilung folgendermaßen definiert,

a. bei n Daten:

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

b. bei klassierten Daten mit absoluten Häufigkeiten:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \cdot n_{ij}$$

c. bei klassierten Daten mit relativen Häufigkeiten:

$$s_{XY} = \frac{n}{n-1} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \cdot h_{ij}$$

Die Kovarianz kennzeichnet das durchschnittliche Abweichungsprodukt der Merkmale X und Y und bildet die Basis der Korrelation. Die Kovarianz lässt die Grundidee der statistischen Korrelation augenscheinlich werden:

- die Gleich- oder
- die Gegenläufigkeit der Abweichungen der jeweiligen Merkmalswerte um ihre Mittelwerte.

Ein großer positiver Wert der Kovarianz ist ein Indiz für eine ausgeprägte positive lineare Korrelation, ein großer negativer Wert der Kovarianz für eine ausgeprägte negative lineare Korrelation.

Allerdings ist die empirische Kovarianz als Korrelationsmaß wenig geeignet, da man für ihre Größe keine Norm kennt. Anders gesagt, die Kovarianz ist betragsmäßig nicht beschränkt. Hinzu kommt noch, dass sie eine dimensionsgeladene Zahl ist, die eine plausible Interpretation erschwert.

standardisierte Kovarianz

Aus diesem Grunde standardisiert man sie mit den empirischen Standardabweichungen s_X und s_Y und interpretiert den Korrelationskoeffizienten als eine **standardisierte Kovarianz**.

2.7 Definition des Korrelationskoeffizienten von BRAVAIS-PEARSON

Nachdem wir die Definition für die Kovarianz formal notiert haben, können wir nun auch die Bestimmung des Korrelationskoeffizienten noch einmal ganz genau festlegen.

Definition

Korrelationskoeffizient nach BRAVAIS-PEARSON

Ist {(x_i, y_i), i = 1,...,n} eine Menge von n Wertepaaren, die an zwei kardinalen Merkmalen X und Y beobachtet wurden, dann heißt die Größe:

$$r_{XY} = \frac{\cancel{\frac{1}{n-1}} \cdot \sum_{i=1}^{n} \overset{\bar{x}_i}{(x_i - \bar{x})} \cdot \overset{\bar{y}_i}{(y_i - \bar{y})}}{\sqrt{\cancel{\frac{1}{n-1}} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\cancel{\frac{1}{n-1}} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{s_{XY}}{s_X \cdot s_Y}$$

$\sqrt{\sum \bar{x}_i^2}$ $\sqrt{\sum \bar{y}_i^2}$

Kovarianz

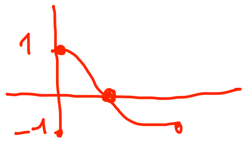
$\approx \cos(\measuredangle(\vec{\bar{x}}, \vec{\bar{y}}))$

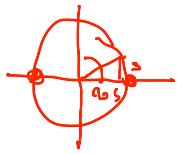
Standardabweichung

Korrelationskoeffizient von X und Y.

Anmerkungen zum Korrelationskoeffizienten

Der Korrelationskoeffizient ist ein normiertes Zusammenhangsmaß.
Der Korrelationskoeffizient misst stets nur





- die Stärke und
- die Richtung (Art)

eines linearen statistischen Zusammenhangs zwischen zwei Merkmalen. Die Betonung liegt auf linear. (Wenn der Zusammenhang zwischen zwei Merkmalen offensichtlich nicht-linear ist, sollte man den Korrelationskoeffizienten nicht ausrechnen).

Der Korrelationskoeffizient r_{XY} kann nur Werte zwischen –1 und +1 annehmen:

$\boxed{-1 \leq r_{XY} \leq 1}$

Die folgende Diashow ist Sir Francis Galton gewidmet, der sich in der zweiten Hälfte des 19. Jahrhunderts mit Fragen der Vererbung beschäftigt hat. Er hat unter anderem beschrieben, dass besonders große oder kleine Eltern zwar große bzw. kleine Nachkommen haben, aber nicht immer extremere.

https://anyflip.com/dkog/ccot

F. GALTON, *Regression towards mediocrity in hereditary stature, Journal of the Anthropological Institute 15 (1886), 246-263.*

Die verwendeten Daten dienen nur der Anschauung. Hier haben wir die Körpergewichte von Vätern und Söhnen angegeben. Sie können die Berechnung des Korrelationskoeffizienten Schritt für Schritt nachvollziehen.

Diashow

Korrelationskoeffizient: Körpergewicht Väter und Söhne

i	x_i	y_i	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	65	68	2.78	0.17	-0.69
2	63	66	13.44	2.51	5.81
3	67	68	0.11	0.17	0.14
4	64	65	7.11	6.67	6.89
5	68	69	1.78	2.01	1.89
6	62	66	21.78	2.51	7.39
7	70	68	11.11	0.17	1.39
8	66	65	0.44	6.67	1.72
9	68	71	1.78	11.67	4.56
10	67	67	0.11	0.34	-0.19
11	69	68	5.44	0.17	0.97
12	71	70	18.78	5.84	10.47
Σ	800	811	84.67	38.92	40.33

Die arithmetischen Mittel:

$$\bar{x} = 66.67 \qquad \bar{y} = 67.58$$

Die Standardabweichungen:

$$s_x = 2.77 \qquad s_y = 1.88$$

Die Kovarianz:

$$s_{xy} = 3.67$$

Der Korrelationskoeffizient:

$$r_{xy} = 0.7$$

2.8 Streudiagramm mit positiver Korrelation

In den folgenden Abschnitten haben Sie die Gelegenheit, sich Richtung und Stärke von Zusammenhängen an Streudiagrammen zu veranschaulichen. Wir zeigen Ihnen zunächst positive Zusammenhänge.

Zur systematischen Betrachtung betrachten wir noch einmal das Streudiagramm. Durch die Mittelwertslinien ist das Diagramm in vier Quadranten aufgeteilt.

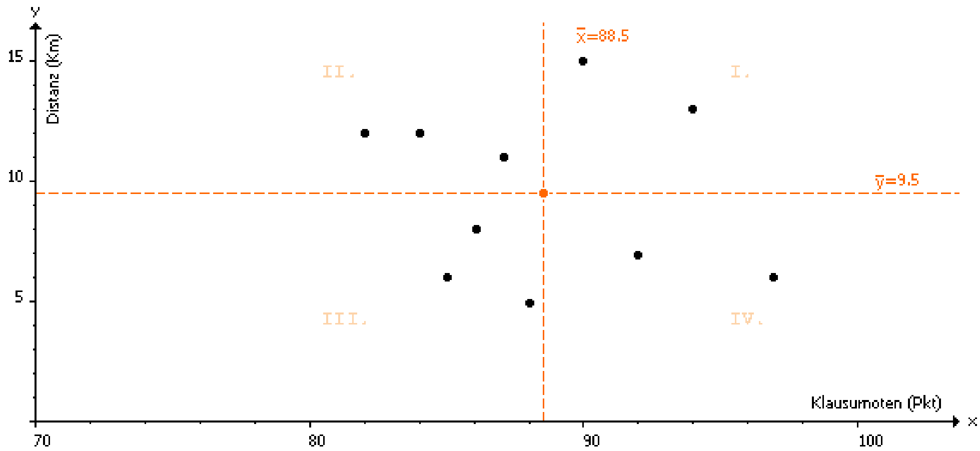


Abb.: Streudiagramm der erreichten Klausurpunkte und der Entfernung des Wohnortes der Studierenden von der Hochschule

Liegen die Punkte im Streudiagramm hauptsächlich in den Quadranten I und III, so liegt eine positive Korrelation vor. Zur Interpretation des Korrelationskoeffizienten eines positiven linearen Zusammenhangs sei gesagt:

- Liegt r_{XY} nahe 0, dann ist das ein Indiz dafür, dass zwischen den Merkmalen X und Y statistisch kein linearer Zusammenhang nachweisbar ist bzw. dass die Merkmale X und Y (linear) voneinander unabhängig sind.
- Ein Wert r_{XY} zwischen 0 und 0,2 kennzeichnet einen sehr schwachen gleichläufigen linearen statistischen Zusammenhang mit einer positiven Steigung.
- Ein Wert r_{XY} zwischen 0,2 und 0,5 kennzeichnet einen schwachen gleichläufigen linearen statistischen Zusammenhang mit einer positiven Steigung.
- Ein Wert r_{XY} zwischen 0,5 und 0,8 kennzeichnet einen mittleren gleichläufigen linearen statistischen Zusammenhang mit einer positiven Steigung.
- Ein Wert r_{XY} zwischen 0,8 und 1 kennzeichnet einen starken gleichläufigen linearen statistischen Zusammenhang mit einer positiven Steigung.
- Gilt r_{XY} = 1, dann liegen alle Wertepaare auf einer Geraden mit positiver Steigung: die Merkmale X und Y sind perfekt linear abhängig.

Die folgende Abbildung zeigt positive Zusammenhänge in Zahlen und Bildern.

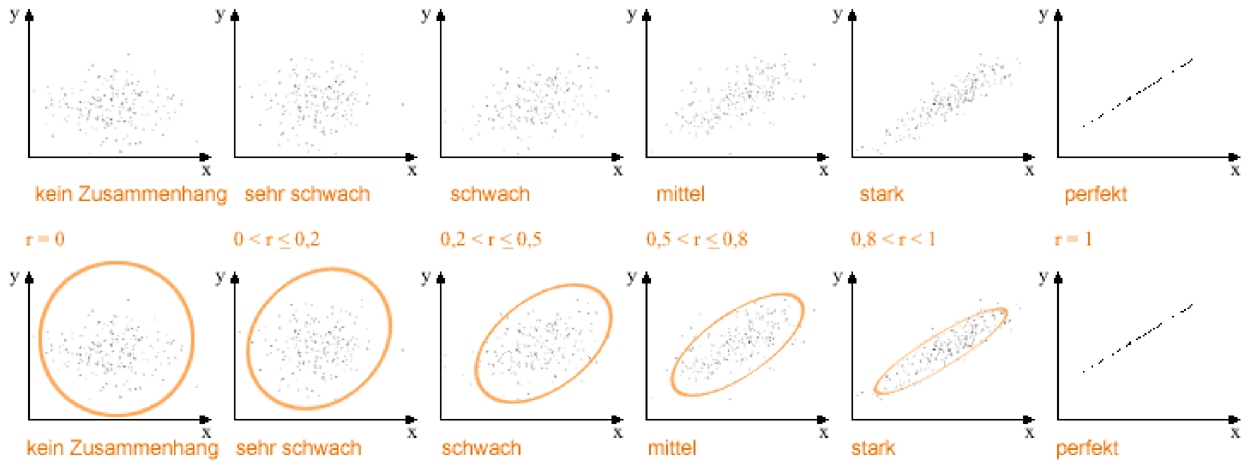


Abb.: Streudiagramme positiver linearer Zusammenhänge verschiedener Stärke mit und ohne die dazugehörigen Hilfskreise und Korrelationskoeffizienten

2.9 Streudiagramm mit negativer Korrelation

Haben Sie Lust auf noch mehr? Hier sind graphische Veranschaulichungen negativer Zusammenhänge.

Liegen die Punkte im Streudiagramm hauptsächlich in den Quadranten II und IV, so liegt eine negative Korrelation vor. Zur **Interpretation** des Korrelationskoeffizienten eines negativen linearen Zusammenhangs sei gesagt:

- Liegt r_{XY} nahe 0, dann ist das ein Indiz dafür, dass zwischen den Merkmalen X und Y statistisch kein linearer Zusammenhang nachweisbar ist bzw. dass die Merkmale X und Y (linear) voneinander unabhängig sind.
- Ein Wert r_{XY} zwischen 0 und – 0,2 kennzeichnet einen **sehr schwachen** gegenläufigen linearen statistischen Zusammenhang mit einer negativen Steigung.
- Ein Wert r_{XY} zwischen – 0.2 und – 0,5 kennzeichnet einen **schwachen** gegenläufigen linearen statistischen Zusammenhang mit einer negativen Steigung.
- Ein Wert r_{XY} zwischen – 0,5 und – 0,8 kennzeichnet einen **mittleren** gegenläufigen linearen statistischen Zusammenhang mit einer negativen Steigung.
- Ein Wert r_{XY} zwischen – 0,8 und – 1 kennzeichnet einen **starken** gegenläufigen linearen statistischen Zusammenhang mit einer negativen Steigung.
- Gilt r_{XY} = – 1, dann liegen alle Wertepaare auf einer Geraden mit negativer Steigung: die Merkmale X und Y sind perfekt linear abhängig.

Die folgende Abbildung zeigt negative Zusammenhänge in Zahlen und Bildern.

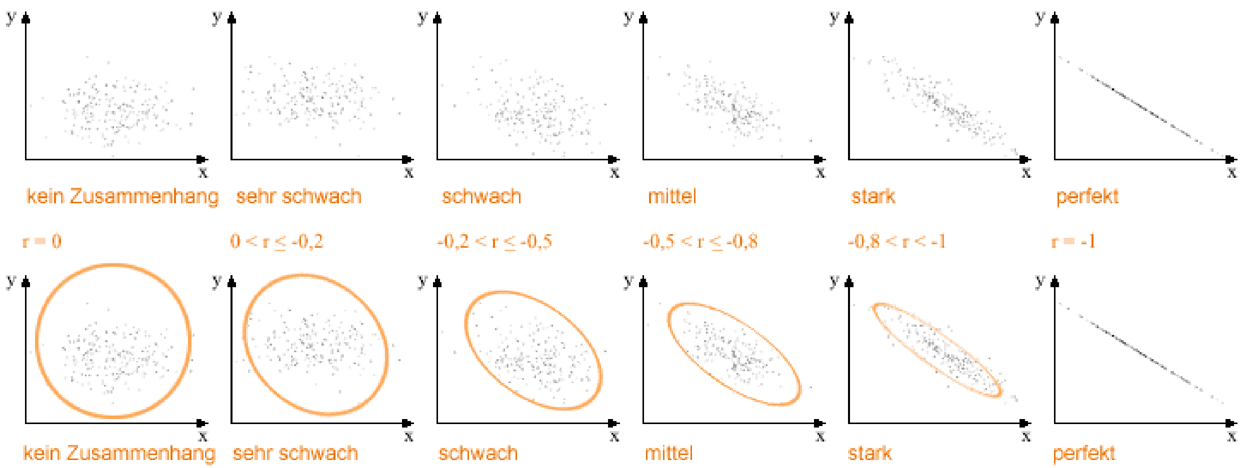


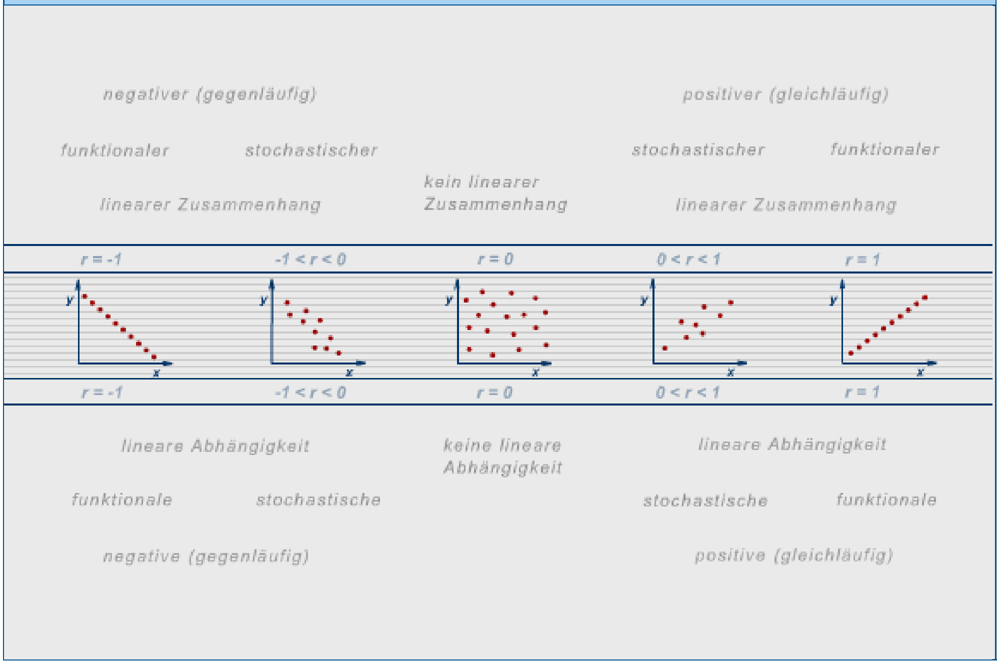
Abb.:
Streudiagramme negativer linearer Zusammenhänge verschiedener Stärke ohne und mit die dazugehörigen Hilfskreise und Korrelationskoeffizienten

2.10 Lineare Zusammenhänge

Die folgende Animation soll Ihnen dabei helfen, Korrelationen aus einem Streudiagramm ohne Berechnung abzuschätzen. Sie können auch versuchen, die Korrelationen zu erraten und dann abzulesen. Aber nicht schummeln, bitte!

Rolloverbild

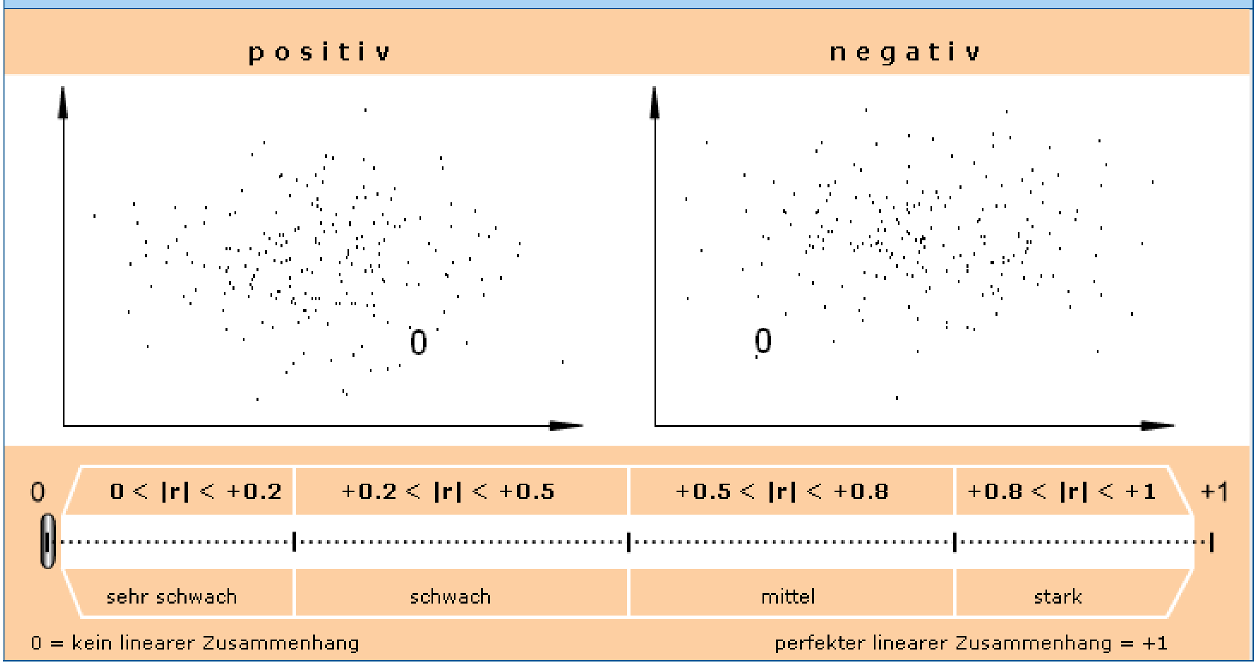
Lineare Zusammenhänge und Abläufe



Die Animation zeigt Streudiagramme positiver und negativer linearer Zusammenhänge verschiedener Stärke mit ihren jeweils dazugehörigen Korrelationskoeffizienten. In der vorliegenden Interaktion lässt sich die Stärke des Zusammenhanges regulieren. Die Stärke des Zusammenhanges wird durch fest definierte Schritte, der Größe |0.01| reguliert.

Interaktion

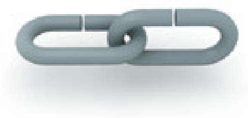
Positive/negative lineare Zusammenhänge und zugehörige Korrelationskoeffizienten



3 Scheinkorrelation

Leider können wir Korrelationen nicht blind vertrauen. Enge Korrelationen können auch künstlich entstehen, ohne dass zwischen den betrachteten Merkmalen ein tatsächlicher Zusammenhang besteht.

Die falsche (sachlich nicht gerechtfertigte) kausale Interpretation gehört zu den bekanntesten Fehlinterpretationen der Korrelation.



Wenn X und Y miteinander korrelieren, so kann dies bedeuten, dass:

- X die Ursache von Y ist,
- Y die Ursache von X ist: Was Ursache und was Wirkung ist, lässt sich wegen der Symmetrie von r_{XY} nicht allein anhand der Korrelation feststellen.
- X und Y rein zufällig in einer entsprechend kleinen Stichprobe miteinander korrelieren, in der Grundgesamtheit jedoch nicht. (Mit solchen Fragen beschäftigt sich die Induktiven Statistik.)
- X und Y nur deshalb miteinander korrelieren, weil sie gemeinsam von einer dritten Variablen Z (Scheinkorrelation) abhängig sind und mit Z (nicht direkt miteinander) in einer Kausalbeziehung stehen.

Wegen dieser Nichteindeutigkeit meint man auch sehr oft, dass Korrelation und Kausalität nichts miteinander zu tun hätten. Das wäre allerdings falsch. Vielmehr sollte man versuchen, Scheinzusammenhänge auszuschließen, wenn Korrelationen verwendet werden.

Achtung

Scheinkorrelation
Sind zwei Variable X und Y nur deshalb hoch korreliert, weil sie gemeinsam von einer dritten Variablen Z abhängig sind, so spricht man von der Scheinkorrelation.

3.1 Anmerkungen zur Scheinkorrelation

Die Korrelation zwischen Storchennestern und Geburten ist das beliebteste Beispiel für eine Scheinkorrelation. Tatsächlich ergibt sich für viele Länder ein positiver Korrelationskoeffizient zwischen der Anzahl der Storchennester und der Geburtenrate über die Zeit. Selbst elementarste Biologiekenntnisse genügen, um an einem Kausalzusammenhang zu zweifeln. Oder?

Die eigentliche Ursache ist leicht gefunden.

Mit der Urbanisierung (die „dahinterstehende" Variable) wurde den Störchen der Lebensraum genommen und mit der wirtschaftlichen Entwicklung wurden die Familiengrößen kleiner.

In diesem Fall liegt sehr offensichtlich nur eine Scheinkorrelation und keine „echte" (d. h. kausal zu interpretierende) Korrelation vor, man spricht auch von „nonsense correlation".

Beispiel

Nonsense Correlation

- Schuhgröße und Intelligenz von Kindern
- Konfession und Körpergröße.

Also, ein grafisch überzeugender Zusammenhang ist kein Beweis für einen Ursache-/Wirkungszusammenhang.

Bei vielen Fällen einer Scheinkorrelation ist es weniger offensichtlich, dass eine Kausalinterpretation nicht zulässig ist. Bei der Korrelation von Zeitreihen, die einen gemeinsamen Trend haben, ist es sehr häufig der Fall. Die trendbereinigten Zeitreihen X' und Y' korrelieren dann weniger miteinander als die noch trendbehafteten Ursprungswerte X und Y. In der Wirtschaftsstatistik tritt das sehr häufig bei der Korrelation mit Sozialproduktsgrößen oder allen wertmäßigen und damit von der Inflation tangierten Größen auf.

Häufig entsteht die Scheinkorrelation auch durch Aggregation von Daten. Bei der Disaggregation zeigt sich, dass sich die Korrelation verringert, d. h. dass sie bei der Bezugnahme auf homogenere Gesamtheiten nicht gilt.

Das Wirken einer dritten Variablen Z geschieht bei der Scheinkorrelation meist nach Art folgender Abbildung.

Abb.: Scheinkorrelation zwischen X und Y

Das Pfeilschema soll andeuten, dass X und Y gemeinsam von Z „verursacht" werden (Fall 1). Hinsichtlich der formalen Zusammenhänge zwischen den Korrelationskoeffizienten sind jedoch die beiden weiteren Fälle der Abbildung nicht unterscheidbar.

Beispiel

Scheinkorrelation

Zwischen der Anzahl der Feuerwehrlöschzüge (X) und der Größe des Brandschadens (Y) besteht eine Korrelation. Der dritte Faktor (Z) ist die Größe des Brandes (z. B. die Flammenmenge).

Die falsche kausale Interpretation würde lauten: Je mehr Feuerwehrlöschzüge bei einem Brand eingesetzt werden, desto größer ist der Brandschaden. Hier wäre also der Feuerwehreinsatz die Ursache des Brandschadens.

3.2 Übung zur Scheinkorrelation

Am folgenden Beispiel können Sie sehen, dass Scheinkorrelationen tatsächlich auftreten. Beim Aufklären des Sachverhalts müssen Sie noch einige Korrelationskoeffizienten berechnen. Danach sollte die Berechnung für Sie wirklich nicht mehr schwierig sein.

Berechnen

Übung KOR-02

Korrelieren Schuhgrößen und Monatseinkommen?

Für das Weinfachgeschäft Maestro seien die folgenden Daten über die Schuhgröße (X) und das Monatseinkommen (Y) getrennt nach Mitarbeiterinnen und Mitarbeiter gegeben (Angaben in €):

Frauen		Männer	
Schuhgröße X	**Monatseinkommen Y**	**Schuhgröße X**	**Monatseinkommen Y**
35	1700	41	2300
36	1400	42	3000
37	1000	43	2700
38	1300	44	2500
39	1100	45	2000

Die Korrelation beträgt r = 0,7.

Heißt dies, dass man deshalb mehr verdient, weil man große Schuhe trägt?

1. Bestimmen Sie die Korrelation zwischen der Schuhgröße und dem Monatseinkommen für alle Wertepaare sowie getrennt für Männer und Frauen. Erklären Sie den Unterschied.

Lösung (Siehe Anhang)

Bearbeitungszeit: 15 Minuten

Zusammenfassung

- Der Korrelationskoeffizient ist ein Maß für Richtung und Stärke des linearen Zusammenhangs zweier Merkmale.
- Die Korrelation ist gleich der mit den Standardabweichungen der Variablen normierten Kovarianz.
- Korrelationen haben Werte zwischen -1 und 1.
- Gilt $r^2 = 1$, dann besteht ein perfekter linearer Zusammenhang.
- Wenn die Korrelation gleich Null ist, besteht kein linearer Zusammenhang. Es kann aber ein anderer Zusammenhang vorhanden sein.
- Hohe Korrelationen bedeuten nicht automatisch eine echte Beziehung, es gibt auch Scheinkorrelationen.
- Korrelationskoeffizienten sollten nur im Streudiagramm interpretiert werden.

Sie sind am Ende dieser Lerneinheit angelangt. Auf der folgenden Seite finden Sie noch die Übungen zur Wissensüberprüfung, weitere Übungen und wichtige Formeln.

Wissensüberprüfung

Multiple Choice

Übung KOR-03

Welche Aussagen sind falsch und welche richtig?

	Richtig	Falsch	Auswertung
Der Korrelationskoeffizient r_{xy} von BRAVAIS und PEARSON ist eine Maßzahl der Stärke und der Richtung (Art) des linearen Zusammenhangs zweier Merkmale.			
Ein Wert r_{xy} zwischen 0.8 und 1 kennzeichnet einen sehr schwachen gegenläufigen statistischen Zusammenhang.			
Ein Wert von r_{xy} nahe 0 ist ein Indiz dafür, dass zwischen den Merkmalen ein gegenläufiger funktionaler Zusammenhang nachweisbar ist bzw. dass die Merkmale linear abhängig sind.			
Ein Wert r_{xy} von -1 ist ein Indiz dafür, dass zwischen den Merkmalen statistisch kein linearer Zusammenhang nachweisbar ist bzw. dass die Merkmale linear unabhängig sind.			

Multiple Choice

Übung KOR-04

Welche Aussagen sind falsch und welche richtig?

	Richtig	Falsch	Auswertung
Ein r_{xy} zwischen 0.2 und 0.5 kennzeichnet einen mittleren gleichläufigen linearen statistischen Zusammenhang.			
Von der Scheinkorrelation spricht man dann, wenn zwei Variablen nur deshalb hoch korreliert sind, weil sie gemeinsam von einer dritten Variable abhängig sind.			
Für einen starken gleichläufigen linearen statistischen Zusammenhang gilt das Folgende: die unter- bzw. überdurchschnittlichen Werte des Merkmals X gehen in der Regel mit den unter- bzw. überdurchschnittlichen Werten des Merkmals Y einher.			

Übungen mit der Statistiksoftware R

Die in der Lerneinheit behandelten Themen können Sie anhand der folgenden Übungsaufgaben mit der Statistiksoftware **R** bearbeiten. Dazu muss die Software „**R**" auf Ihrem Rechner installiert sein.

Installationshinweise [Manuals | **R** Installation and Administration]

Sie können die R-Datei mit der Lösung auf Ihrem Rechner speichern und die Datei mit **R**. öffen. Copy und Paste sollte auch funktionieren. Nur beim Einlesen von Daten aus einer Datei muss diese Datei im Arbeitsverzeichnis von R gespeichert sein.

Für einige Übungen stehen auch Musterlösungen für das Programm Excel bereit.

Berechnen

Übung KOR-05a

Spiroergometrie

In der Datei **`spiro.txt`** finden Sie die Ergebnisse eines Leistungstests auf dem Laufband, bei dem die Herzfrequenz (HF) sowie die relative Sauerstoffaufnahme (in ml/min/kg) (VO2) bei steigender Geschwindigkeit (V) gemessen wurde.

`spiro.txt` (2 KB)

Aufgaben

1. Berechnen Sie alle möglichen Korrelationen der drei Variablen V, HF und VO2 und beurteilen Sie die Ergebnisse.
2. Stellen Sie die Variablen mit der höchsten Korrelation in einem Streudiagramm grafisch dar.

Lösung mit R und Excel (Siehe Anhang)

Bearbeitungszeit: 20 Minuten

Berechnen

Übung KOR-05b

Katholiken

Die folgende Tabelle enthält die durchschnittliche Körpergröße und den Prozentsatz an Katholiken in verschiedenen europäischen Staaten.

i	Staat	Größe (cm)	Anteil Katholiken in %
1	Belgien	169,2	80,59
2	Dänemark	172,4	0,66
3	Spanien	166,4	93,51
4	Frankreich	168,9	79,75
5	Irland	169,1	75,68
6	Italien	168,0	97,03
7	Niederlande	173,8	34,66
8	Österreich	171,4	75,34
9	Portugal	164,7	92,81
10	Schweden	172,2	1,81

Aufgabe

1. Untersuchen Sie den Datensatz auf einen Zusammenhang zwischen Zugehörigkeit zur katholischen Kirche und Körpergröße!

Lösung mit R und Excel (Siehe Anhang)

Bearbeitungszeit: 20 Minuten

Berechnen

Übung KOR-05c

Wörter merken

Wir wiederholen hier das Beispiel aus der Lerneinheit um Ihnen zu zeigen, wie einfach die Berechnung mit R funktioniert.

In der folgenden Tabelle sind die Ergebnisse eines Tests zur Gedächtnisleistung festgehalten. Es haben insgesamt 9 Studentinnen teilgenommen.

Anzahl auswendig zu lernender Wörter	5	10	15	20	25	30	35	40	45
Prozentsatz der davon behaltenen Wörter (%)	80	70	33	20	20	27	17	23	13

Aufgabe

1. Untersuchen Sie ob zwischen den beiden Merkmalen ein Zusammenhang besteht und welche Stärke und Richtung er besitzt. Lesen Sie die Werte ein, lassen Sie sich die Korrelation und das Streudiagramm ausgeben.

Lösung mit R (Siehe Anhang)

Bearbeitungszeit: 20 Minuten

Zusätzliche Übungsaufgaben

Berechnen

Übung KOR-06

Wettlauf

Die folgende Tabelle enthält die Körpergröße und die Platzierung von 10 Studierenden, die bei einem Sportfest an einem Wettlauf teilgenommen haben:

Studierende	A	B	C	D	E	F	G	H	I	J
Körpergröße	180	170	174	190	165	182	178	169	184	189
Platz	3	7	8	2	10	5	6	9	1	4

Untersuchen Sie den linearen Zusammenhang zwischen der Körpergröße und der Platzierung.

Interpretieren Sie kurz Ihre Ergebnisse.

Lösung (Siehe Anhang)

Bearbeitungszeit: 15 Minuten

Berechnen

Übung KOR-07

Daten

Bestimmen Sie die Korrelation zwischen den Merkmalen X und Y für die folgenden Daten:

x_i	80	79	77	76	74	73	71	69	68	66
y_i	11,2	10,9	11	10,7	10,9	10,7	10,5	10,6	10,3	10,2

Interpretieren Sie kurz Ihr Ergebnis.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung KOR-08

Sonnenscheindauer

In der folgenden Tabelle ist die monatliche Sonnenscheindauer in Stunden am Vormittag
und Nachmittag aufgelistet.

Monat	1	2	3	4	5	6	7	8	9	10	11	12
Vormittag	26	45	111	92	119	114	136	156	132	55	30	35
Nachmittag	36	59	102	90	97	116	114	143	131	59	41	37

Bestimmen Sie die Korrelation zwischen Sonnenscheindauer am Vormittag und Nachmittag. Interpretieren Sie kurz das Ergebnis.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung KOR-09

Körpergröße Sohn und Tochter

Bei einer Umfrage von Familien mit 2 Kindern unterschiedlichen Geschlechtes im Alter bis 24 Monaten wurden folgende Körpergrößen (in cm) von Sohn und Tochter notiert:

Sohn	73, 70, 74, 68, 70, 67, 71, 70, 68, 69, 68, 71, 73, 69, 68
Tochter	69, 67, 63, 66, 67, 64, 68, 67, 65, 65, 61, 66, 67, 66, 67

Untersuchen Sie den linearen Zusammenhang zwischen der Körpergröße des Sohnes und der Tochter. Interpretieren Sie kurz Ihr Ergebnis.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung KOR-10

Handelskette

Untersuchen Sie für 10 Filialen einer Lebensmittel-Handelskette, welcher lineare Zusammenhang zwischen Umsatz (in Mio. Euro) und Verkaufsfläche (in m^2) besteht:

Umsatz	3, 8, 19, 22, 31, 42, 48, 52, 54, 61
Verkaufsfläche	150, 180, 420, 480, 660, 1000, 1300, 1500, 1600, 1710

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Appendix

Lösung für Übung KOR-01

Studierende und Entfernung zur Hochschule

Mittelwerte

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{885}{10} = 88{,}5\ Punkte$$

$$\overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{95}{10} = 9{,}5\ km$$

Zwischenergebnisse

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	88	5	- 0,5	- 4,5	2,25	0,25	20,25
2	94	13	5,5	3,5	19,25	30,25	12,25
3	85	6	- 3,5	- 3,5	12,25	12,25	12,25
4	86	8	- 2,5	- 1,5	3,75	6,25	2,25
5	82	12	- 6,5	2,5	- 16,25	42,25	6,25
6	97	6	8,5	- 3,5	- 29,75	72,25	12,25
7	92	7	3,5	- 2,5	- 8,75	12,25	6,25
8	87	11	- 1,5	1,5	- 2,25	2,25	2,25
9	84	12	- 4,5	2,5	- 11,25	20,25	6,25
10	90	15	1,5	5,5	8,25	2,25	30,25
∑	**885**	**95**	**0**	**0**	**- 22,5**	**200,5**	**110,5**

Streudiagramm (Interpretation)

Die Punktewolke lässt (etwa im Unterschied zu den bisher betrachteten Punktewolken) keinen linearen Zusammenhang zwischen der Punktzahl und der Entfernung des Wohnortes erkennen.

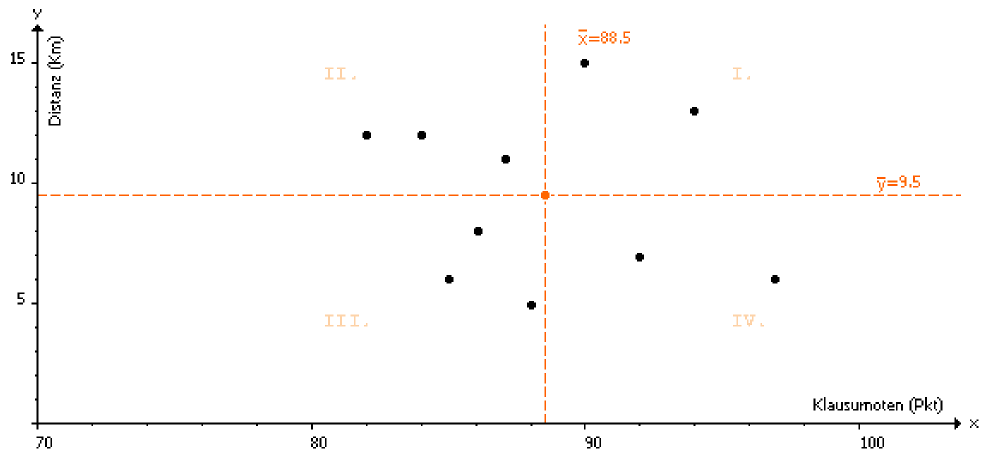


Abb.: Streudiagramm der erreichten Punkte in einer Statistikklausur und der Entfernung des Wohnortes der Studierenden von der Hochschule

Korrelationskoeffizient (Interpretation)

Die nahezu kreisförmige Punktewolke entspricht dem zugehörigen Korrelationskoeffizienten:

$$r_{XY} = \frac{s_{XY}}{s_x s_Y} = \frac{-2{,}5\ Pkt\ \cdot\ km}{4{,}7\ Pkt\ \cdot\ 3{,}5\ km} \approx -0{,}15$$

Empirische Kovarianz s_{XY}:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{-22{,}5}{9} = -2{,}5\ Pkt \cdot km$$

Empirische Standardabweichungen:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{200{,}5}{9}} = 4{,}7\ Pkt$$

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{110{,}5}{9}} = 3{,}5\ km$$

Der Wert des Korrelationskoeffizienten nimmt einen Wert nahe Null an.
Zwischen den Punkten der Statistikklausur und der Entfernung des Wohnortes der n = 10 Studierenden zur Hochschule besteht ein vernachlässigbarer linearer Zusammenhang. Anhand des vorliegenden statistischen Befundes kann davon ausgegangen werden, dass die Klausurnoten und die Entfernung des Wohnortes zur Hochschule nicht miteinander korrelieren. Jedes andere Ergebnis hätte uns zumindest verwundert.

Es ist ohne Belang, ob man im konkreten Fall den Zusammenhang zwischen den Punkten einer Klausur und der Entfernung des Wohnortes zur Hochschule oder den Zusammenhang zwischen der Entfernung des Wohnortes zur Hochschule und der Klausurnoten statistisch analysiert, weil der einfache lineare Korrelationskoeffizient ein symmetrisches **Zusammenhangsmaß** ist.

Wollte man zum Beispiel versuchen, die Klausurnoten allein aus der Entfernung vorherzusagen hätte man es mit einem nicht symmetrischen, gerichteten Zusammenhang zu tun. Diese Form der statistischen Analyse fasst man unter dem Begriff der Regressionsanalyse zusammen. Die Regressionsanalyse ist ein spezieller Gegenstand, der Ihnen in Lerneinheit *„ELR – ‚Einfache lineare Regression'"* vorgestellt wird.

Lösung für Übung KOR-02

Korrelieren Schuhgrößen und Monatseinkommen?

Es liegt ein typischer Fall von Scheinkorrelation vor.

Angenommen, bei den ersten fünf Personen handelt es sich um Frauen, die in der Regel eine kleinere Schuhgröße haben als Männer und häufig auch weniger verdienen. (Gerecht ist das nicht.) Die nächsten fünf Personen seien Männer.

Die nachfolgende Tabelle beinhaltet die für die angestrebte Korrelationsanalyse der ersten fünf Personen (Frauen) erforderlichen Zwischenergebnisse.

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	35	1,7	- 2	- 0,4	- 0,8	4	0,16
2	36	1,4	- 1	0,1	- 0,1	1	0,01
3	37	1,0	0	- 0,3	0	0	0,09
4	38	1,3	1	0	0	1	0
5	39	1,1	2	- 0,2	- 0,4	4	0,04
∑	185	6,5	0	0	- 1,3	10	0,3

Der Korrelationskoeffizient nach BRAVAIS und PEARSON der ersten fünf Personen (Frauen) ist:

$$r_{XY} = \frac{s_{XY}}{s_x s_Y} = \frac{-0{,}325 \text{ Schuhgröße } \cdot \text{ Tsd. €}}{1{,}58 \text{ Schuhgröße } \cdot \ 0{,}27 \text{ Tsd. €}} \approx -0{,}75$$

mit empirischer Kovarianz s_{XY}:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{-1{,}3}{4} = -0{,}325 \text{ Schuhgröße} \cdot \text{Tsd. €}$$

und empirischen Standardabweichungen:

$$s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2} = \sqrt{\frac{10}{4}} = 1{,}58 \text{ Schuhgröße}$$

$$s_Y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2} = \sqrt{\frac{0.3}{4}} = 0{,}27 \text{ Tsd. €}$$

Die folgende Tabelle beinhaltet die für die angestrebte Korrelationsanalyse der nächsten fünf Personen (Männer) erforderlichen Zwischenergebnisse.

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	41	2,3	- 2	- 0,2	0,4	4	0,04
2	42	3,0	- 1	0,5	- 0,5	1	0,25
3	43	2,7	0	0,2	0	0	0,04
4	44	2,5	1	0	0	1	0
5	45	2,0	2	- 0,5	- 1	4	0,25
∑	215	12,5	0	0	- 1,1	10	0,58

Der Korrelationskoeffizient nach BRAVAIS und PEARSON der nächsten fünf Personen (Männer) ist:

$$r_{XY} = \frac{s_{XY}}{s_x s_Y} = \frac{-0{,}275 \text{ Schuhgröße } \cdot \text{ Tsd. €}}{1{,}58 \text{ Schuhgröße } \cdot \ 0{,}38 \text{ Tsd. €}} \approx -0{,}46$$

mit empirischer Kovarianz s_{XY}:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{-1{,}3}{4} = -0{,}275 \text{ Schuhgröße} \cdot \text{Tsd.€}$$

und empirischen Standardabweichungen:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{10}{4}} = 1{,}58 \text{ Schuhgröße}$$

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{0{,}58}{4}} = 0{,}38 \text{ Tsd.€}$$

Die folgende Tabelle beinhaltet die für die angestrebte Korrelationsanalyse für alle Personen (Frauen und Männer) erforderlichen Zwischenergebnisse:

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	35	1,7	- 5	- 0,2	1	25	0,04
2	36	1,4	- 4	- 0,5	2	16	0,25
3	37	1,0	- 3	- 0,9	2,7	9	0,81
4	38	1,3	- 2	- 0,6	1,2	4	0,36
5	39	1,1	- 1	- 0,8	0,8	1	0,64
6	41	2,3	1	0,4	0,4	1	0,16
7	42	3,0	2	1,1	2,2	4	1,21
8	43	2,7	3	0,8	2,4	9	0,64
9	44	2,5	4	0,6	2,4	16	0,36
10	45	2,0	5	0,1	0,5	25	0,01
∑	**400**	**19**	**0**	**0**	**15,6**	**110**	**4,48**

Der Korrelationskoeffizient nach BRAVAIS und PEARSON für alle Personen (Frauen und Männer) ist:

$$r_{XY} = \frac{s_{XY}}{s_x s_Y} = \frac{1{,}73 \text{ Schuhgröße } \cdot \text{ Tsd. €}}{3{,}50 \text{ Schuhgröße } \cdot \ 0{,}71 \text{ Tsd. €}} \approx 0{,}70$$

mit empirischer Kovarianz s_{XY}:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{15{,}6}{9} = 1{,}73 \text{ Schuhgröße} \cdot \text{Tsd.€}$$

und empirischen Standardabweichungen:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{110}{9}} = 3{,}50 \text{ Schuhgröße}$$

$$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{4{,}48}{9}} = 0{,}71 \text{ Tsd.€}$$

Wie wir schon gesehen haben, erhält man für die ersten fünf Personen (also für die Frauen) für die Korrelation zwischen X und Y r_{XY} = -0,75 und für die nächsten fünf Personen (also die Männer) r_{XY} = -0,46, bei den beiden Gruppen zusammen aber r_{XY} = 0,70. Man beachte auch, dass sich das Vorzeichen ändert!

Lösung Übung KOR-05a

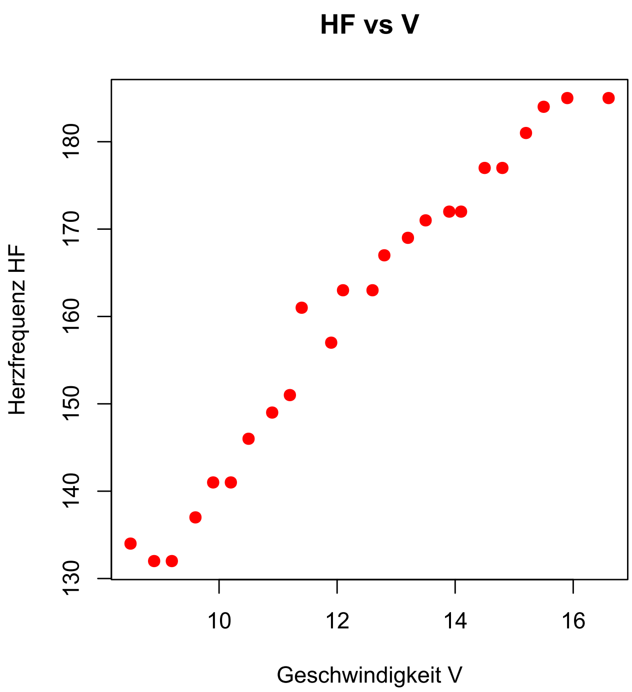
Spiroergometrie

1. Um die Korrelation zwischen den drei Variablen Geschwindigkeit, Herzfrequenz und relative Sauerstoffaufnahme zu berechnen, wird die Funktion **cor** verwendet.

	V	HF	VO2
V	1,0000000	0,9851458	0,9678757
HF	0,9851458	1,0000000	0,9668551
VO2	0,9678757	0,9668551	1,0000000

Beurteilung
Auf der Diagonalen stehen immer Einsen, denn eine Variable korreliert mit sich selbst zu 100 Prozent. Auch die anderen Variablen sind hier stark korreliert da alle Korrelationen einen Wert größer 0.96 aufweisen. Das heißt, mit zunehmender Geschwindigkeit steigt sowohl die Herzfrequenz als auch die relative Sauerstoffaufnahme.

2. Die höchste Korrelation und somit der größte lineare Zusammenhang herrscht zwischen den Variablen Geschwindigkeit (V) und Herzfrequenz (HF). Deshalb zeichnen wir diese in ein Streudiagramm.



Lösung mit R

spiro_loesung.R

```r
# Einlesen der Werte aus der Datei "spiro.txt"
spiro<-read.table("spiro.txt",sep="\t",header=TRUE)


# Ausgabe der Korrelationen
cor(spiro)

# Variablen Geschwindigkeit (V) und Herzfrequenz (HF) als Streudiagramm .
plot(
  spiro$V,
  spiro$HF,
  main = "HF vs V",
  xlab = "Geschwindigkeit V",
  ylab = "Herzfrequenz HF",
  col = "red",
  pch = 16,
  cex = 1.2
)
```

Lösung mit Excel

WMS_KOR_05_Spiroergometrie.xlsx (12 KB)

Aufgabe 1*: Berechnen Sie alle möglichen Korrelationen der drei Variablen V, HF und VO2 und beurteilen Sie die Ergebnisse.*

Um die Korrelation zwischen den drei Variablen Geschwindigkeit, Herzfrequenz und relative Sauerstoffaufnahme zu berechnen, wird die Funktion `KORREL` verwendet. In den einzelnen Zellen muss somit folgendes stehen:

38		V	HF	VO2
39	V	=KORREL(A2:A25;A2:A25)	=KORREL(A2:A25;B2:B25)	=KORREL(A2:A25;C2:C25)
40	HF	=KORREL(B2:B25;A2:A25)	=KORREL(B2:B25;B2:B25)	=KORREL(B2:B25;C2:C25)
41	VO2	=KORREL(C2:C25;A2:A25)	=KORREL(C2:C25;B2:B25)	=KORREL(C2:C25;C2:C25)

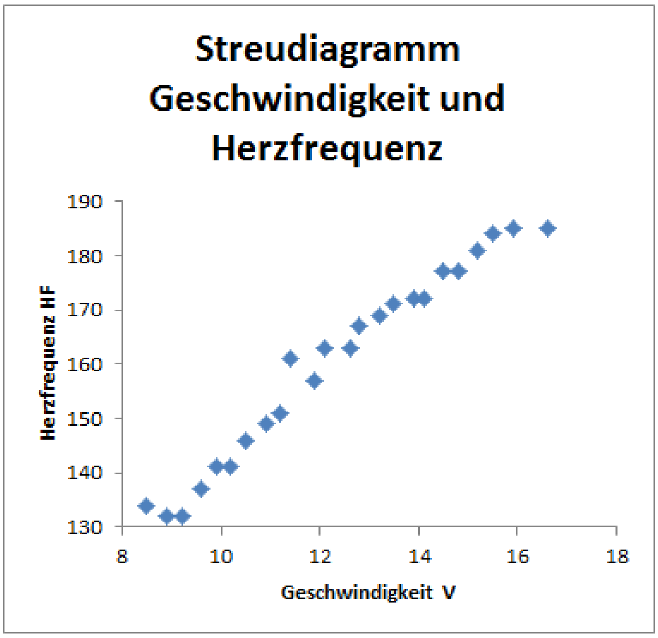
Damit erhalten wir folgende Tabelle:

	V	HF	VO2
V	1,00	0,99	0,97
HF	0,99	1,00	0,97
VO	0,97	0,97	1,00

Auf der Diagonalen stehen immer Einsen, denn eine Variable korreliert mit sich selbst zu 100 Prozent. Auch die anderen Variablen sind hier stark korreliert. Das heißt, mit zunehmender Geschwindigkeit steigt sowohl die Herzfrequenz als auch die relative Sauerstoffaufnahme.

Aufgabe 2*: Stellen Sie die Variablen mit der höchsten Korrelation in einem Streudiagramm grafisch dar.*

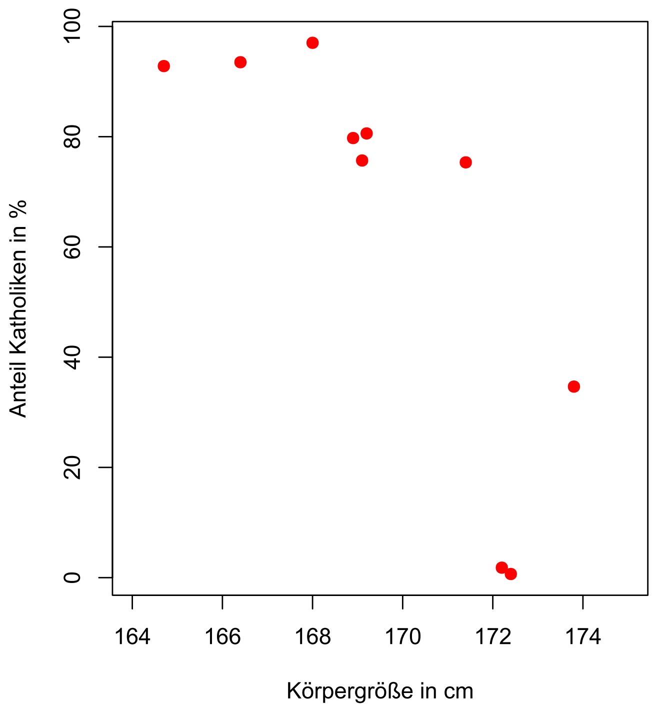
Die stärkste Korrelation zeigen Geschwindigkeit und Herzfrequenz. Deshalb zeichnen wir diese mit Hilfe des Punkt (XY)-Diagramms in ein Streudiagramm.


Streudiagramm
Geschwindigkeit und
Herzfrequenz
Herzfrequenz HF
190
180
170
160
150
140
130
8
10
12
14
16
18
Geschwindigkeit V

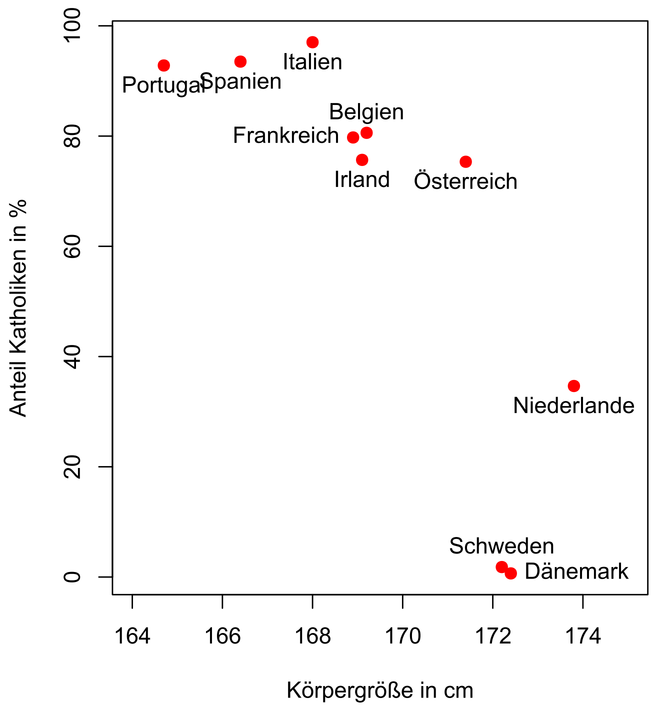
Lösung Übung: KOR-05b

Katholiken

1. Das Streudiagramm und die Korrelation von etwa -0,8 zeigt statistisch gesehen einen negativen linearen Zusammenhang. Das würde bedeuten, dass mit zunehmender durchschnittlicher Körpergröße der Anteil an Katholiken abnimmt. Ob das eine sinnvolle Erkenntnis ist, bleibt dahingestellt.



Die Anzeige der Ländernamen in der Graphik liefert einen Ansatz zur Erklärung.



Lösung mit R

katholiken_loesung.R

```
# Einlesen der Daten
daten <-
  data.frame(
    Staat = c(
      "Belgien",
      "Dänemark",
      "Spanien",
      "Frankreich",
      "Irland",
      "Italien",
      "Niederlande",
      "Österreich",
      "Portugal",
      "Schweden"
    )
    ,
    Groesse = c(169.2, 172.4, 166.4, 168.9, 169.1,
                168, 173.8, 171.4, 164.7, 172.2)
    ,
    Katholiken = c(80.59, 0.66, 93.51, 79.75, 75.68,
                   97.03, 34.66, 75.34, 92.81, 1.81)
  )


#  Streudiagramm
opar<-par(mar=c(5,4,1,1)+.25)
plot(
  daten$Groesse,
  daten$Katholiken,
  pch = 16,
  cex = 1.2,
  col = "red",
  xlab = "Körpergröße in cm",
  ylab = "Anteil Katholiken in %",
  xlim = c(164, 175)
)

#  Korrelation
cor(daten$Groesse,daten$Katholiken)

#  Die Bezeichnung der Länder in der Graphik liefert einen Ansatz zur Erklärung
for(i in 1:10)
{
  text(daten$Groesse[i],
       daten$Katholiken[i],
       daten$Staat[i],
       pos = position[i])
}

# par setzt die Anzeige der Ländernamen wieder zurück.
par(opar)
```

Lösung mit Excel

WMS_KOR_05_Katholiken.xlsx (10 KB)

***Aufgabe 1**: Untersuchen Sie den Datensatz auf einen Zusammenhang zwischen Zugehörigkeit zur katholischen Kirche und Körpergröße!*

Um herauszufinden, ob ein Zusammenhang zwischen durchschnittlicher Körpergröße der Menschen eines europäischen Staates und der Anzahl der in ihm lebenden Katholiken besteht, berechnen wir die Kovarianz und den Korrelationskoeffizient. Dafür verwenden wir die Funktionen `KOVARIANZ.S` und `KORREL`.

20	Kovarianz:
21	=KOVARIANZ.S(C2:C11;D2:D11)
22	
23	Korrelation:
24	=KORREL(C2:C11;D2:D11)

Kovarianz:
-83,75
Korrelation:
-0,79

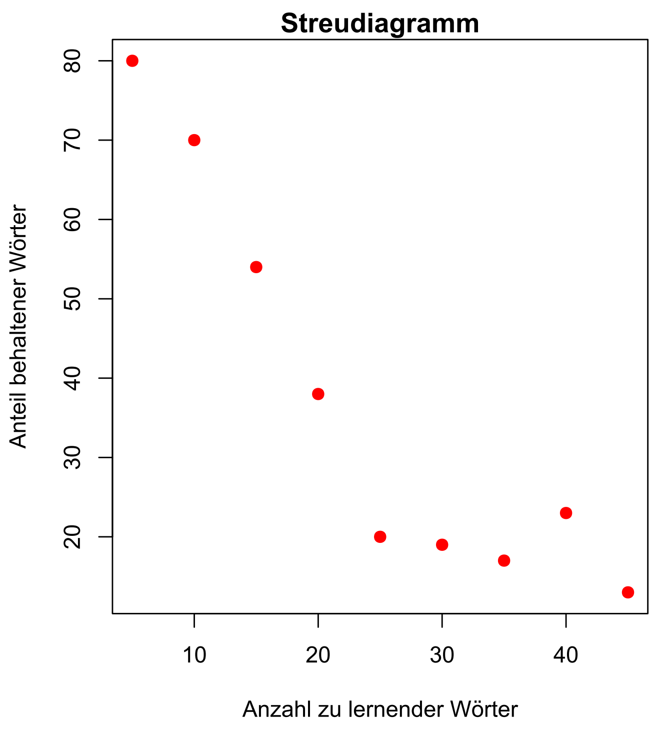
Die Kovarianz ist negativ und der Korrelationskoeffizient weist auf eine fast starke Korrelation zwischen durchschnittlicher Körpergröße und Anzahl der Katholiken hin. Das hieße, je kleiner die durchschnittliche Körpergröße ist, desto größer ist die Anzahl der Katholiken in einem Staat. Da dieses Ergebnis aber keinen Sinn macht, ist davon auszugehen, dass es sich hierbei um eine Scheinkorrelation handelt.

Lösung für Übung KOR-05c

Wörter merken

Die Korrelation zwischen Anzahl der Worte und den gemerkten Worten beträgt -0,9153434

Am Streudiagramm und an der Korrelation ist zu erkennen, dass ein negativer linearer Zusammenhang herrscht! Je mehr Wörter zu lernen waren, desto kleiner war der Anteil der behaltenen Wörter.



Lösung mit R

woerter_loesung.R

```r
# Einlesen der Daten
daten <-
  data.frame(
    X = c(5, 10, 15, 20, 25, 30, 35, 40, 45),
    Y = c(80, 70, 54, 38, 20, 19, 17, 23, 13)
  )

#  Zeichnen eines Streudiagramms
plot(
  daten$X,
  daten$Y,
  pch = 16,
  cex = 1.2,
  col = "red",
  main = "Streudiagramm",
  xlab = "Anzahl zu lernender Wörter",
  ylab = "Anteil behaltener Wörter"
)

#  Berechnung der Korrelation
cor(daten$X,daten$Y)
```

Lösung mit Excel

Lösung für Übung KOR-06

Wettlauf

$$\overline{x} = \frac{1781}{10} = 178.1\ ,\ \overline{y} = \frac{55}{10} = 5.5$$

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
180	3	1,9	- 2,5	- 4,75	3,61	6,25
170	7	- 8,1	1,5	- 12,15	65,61	2,25
174	8	- 4,1	2,5	- 10,25	16,81	6,25
190	2	11,9	- 3,5	- 41,65	141,61	12,25
165	10	- 13,1	4,5	- 58,95	171,61	20,25
182	5	3,9	- 0,5	- 1,95	15,21	0,25
178	6	- 0,1	0,5	- 0,5	0,01	0,25
169	9	- 9,1	3,5	- 31,85	82,81	12,25
184	1	5,9	- 4,5	- 26,55	34,81	20,25
189	4	10,9	- 1,5	- 16,35	118,81	2,25
Σ		**0**	**0**	**- 204,50**	**650,90**	**82,50**

Der Korrelationskoefizient nach BRAVAIS und PEARSON ist:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-22{,}72}{\sqrt{72{,}32}\sqrt{9{,}16}} = -0{,}8824$$

Empirische Kovarianz s_{xy}:

$$s_{xy} = \frac{-1}{9} \cdot 204{,}5 = -22{,}72$$

Empirische Standardabweichungen:

$$s_x^2 = \frac{1}{9} \cdot 650{,}90 = 72{,}32$$

$$s_y^2 = \frac{1}{9} \cdot 82{,}50 = 9{,}16$$

Interpretation
Die Daten sind hoch negativ korreliert. Man kann sagen, dass je größer die Person ist, desto besser die Platzierung.

Lösung für Übung KOR-07

Daten

$$\overline{x} = \frac{733}{10} = 73{,}3 \text{ , } \overline{y} = \frac{107}{10} = 10{,}7$$

$$s_x = \sqrt{\frac{204{,}10}{9}} = \sqrt{20{,}41} = 4{,}762 \text{ , } s_y = \sqrt{\frac{0{,}88}{9}} = \sqrt{0{,}088} = 0{,}313$$

$$s_{xy} = \frac{12{,}40}{9} = 1{,}378$$

$$r = \frac{1{,}378}{4{,}762 \cdot 0{,}313} = 0{,}93$$

Es liegt mit r = 0,93 ein sehr starker positiver Zusammenhang vor.

Lösung für Übung KOR-08

Sonnenscheindauer

$$\overline{x} = 87{,}58 \text{ , } \overline{y} = 85{,}42$$

$$s_x = 46{,}71 \text{ , } s_y = 37{,}82$$

$$s_{xy} = 1732{,}83 \text{ , } r = \frac{1732{,}83}{1766{,}57} = 0{,}98$$

Interpretation: Zwischen Sonnenscheindauer am Vormittag und Nachmittag besteht ein starker positiver Zusammenhang.

Lösung für Übung KOR-09

Körpergröße Sohn und Tochter

$\overline{x} = 69{,}93$, $\overline{y} = 65{,}87$

$s_x = 2{,}12$, $s_y = 2{,}03$

$s_{xy} = 1{,}35$, $r = \dfrac{1{,}35}{4{,}3} = 0{,}31$

Interpretation: Zwischen der Körpergröße des Sohnes und der Tochter besteht kein bedeutender Zusammenhang.

Lösung für Übung KOR-10

Handelskette

$\overline{x} = 34$, $\overline{y} = 900$

$s_x = 20{,}35$, $s_y = 598{,}28$

$s_{xy} = 12000$, $r = \dfrac{12000}{12174{,}998} = 0{,}9856$

Es liegt mit r = 0,9856 ein sehr starker positiver Zusammenhang vor.