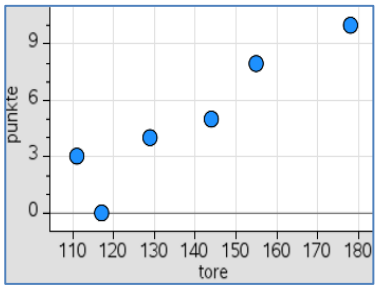
Lineare Regression Teil 2 – Der Korrelationskoeffizient

Die unten stehende Tabelle zeigt den Endstand in der Hauptrundengruppe I bei der Handballweltmeisterschaft 2025. Stellt man beispielsweise den Zusammenhang zwischen den erzielten Toren und den erreichten Punkten grafisch dar, so ergibt sich das danebenstehende Diagramm[1].

HAUPTRUNDE GRUPPE I

Platz	Team	Sp.	S	U	N	Tore	Diff.	Pkt.
1	Dänemark	5	5	0	0	178:121	57	10
2	Deutschland	5	4	0	1	155:137	18	8
3	Schweiz	5	2	1	2	144:138	6	5
4	Italien	5	2	0	3	129:149	-20	4
5	Tschechien	5	1	1	3	111:125	-14	3
6	Tunesien	5	0	0	5	117:164	-47	0



Die Behauptung „Je mehr Tore erzielt werden, desto höher ist die erreichten Punktzahl.“ stimmt sicher so nicht, denn der Tabellenletzte hat hier mehr Tore erzielt als der Tabellenvorletzte, aber weniger Punkte erhalten. Der Formulierung „Je mehr Tore erzielt werden, desto größer ist in der Regel eine hohe Punktzahl.“ könnte man vermutlich eher zustimmen.

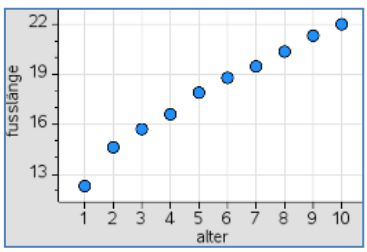
Mathematiker haben eine Methode entwickelt, die die Qualität solcher Zusammenhänge zwischen zwei Variablen durch eine Zahl, den „Korrelationskoeffizienten“ beschreibt.

Im Folgenden wird zunächst beschrieben, wie diese Zahl definiert wird. Dann geht es darum zu zeigen, wie der Korrelationskoeffizient mit einem einfachen wissenschaftlichen Taschenrechner, dem TI-30X Prio MathPrint™, einem der durch die Kultusministerkonferenz für das Abitur ab 2029 zugelassenen Taschenrechner, ermittelt werden kann.

Korrelation beschreibt den Zusammenhang zweier Variablen

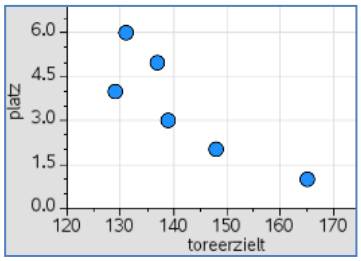
Beispiele

Zwei Variablen werden als **„positiv korreliert“** beschrieben, wenn große Werte der einen Variablen im Allgemeinen mit großen Werten der anderen und kleine Werte der einen mit kleinen Werten der anderen zusammengehen.
(Je älter ein Kind wird, desto größer ist seine Fußlänge.)

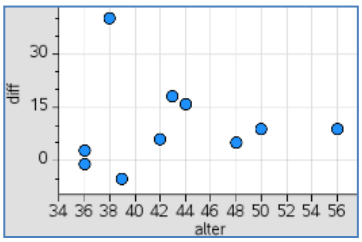


[1] Alle Diagramme wurden mit der Software von TI-Nspire erstellt.

Hingegen heißen zwei Variablen **„negativ korreliert“**, wenn im Allgemeinen große Werte der einen mit kleinen Werten der anderen und kleine Werte der einen mit großen Werten der anderen Variablen zusammengehen.
(Je mehr Tore eine Mannschaft während eines Wettbewerbs erzielt, desto besser ist ihre Platzierung, also desto kleiner ist ihre Platznummer.)



Zwei Variablen werden als **„unkorreliert“** bezeichnet, wenn große Werte der einen sowohl mit großen als auch mit kleinen Werten der anderen und umgekehrt zusammen auftreten.
(Das Alter der Trainer korreliert nicht mit der Tordifferenz ihrer Mannschaften bei einem Fußballturnier.)

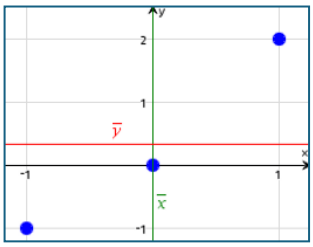


Erläuterung
Was aber heißt „groß“ und was heißt „klein“ in solchen Zusammenhängen? Dazu schreibt der bekannte Statistikprofessor Walter Krämer (s. Krämer, S. 110 – 112) u.a.: „Das konkrete Ausmaß der Korrelation und dessen Messung ist ein Problem für sich. Seine Lösung geht auf den Engländer Sir Francis Galton (1822-1911) zurück. […] Die Lösung geschieht in zwei Etappen. Der erste Schritt ist, zu entscheiden: Was heißt „groß“ und was heißt „klein“? […] Galton hat diesen gordischen Knoten elegant durchgeschnitten: „groß“ ist größer als der Durchschnitt und „klein“ ist kleiner als der Durchschnitt. Der zweite Schritt besteht darin, die jeweiligen Abweichungen vom Durchschnitt geeignet zu gewichten. Angenommen, ein Objekt […] ist in beiden Variablen größer als der Durchschnitt. Dann gehen die Abweichungen vom Durchschnitt je nach Ausmaß unterschiedlich in den Korrelationskoeffizienten ein. Wenn ein Mann bei Größe und Gewicht nur knapp den Durchschnitt überschreitet, so spricht das weniger für eine positive Korrelation, als wenn er bei beiden Variablen den Durchschnitt beträchtlich überschreitet (und spiegelbildlich, wenn er bei beiden Variablen den Durchschnitt beträchtlich unterschreitet). Liegt er gar bei der Größe über und beim Gewicht unter dem Durchschnitt, ist das sogar ein Gegenargument. Galtons Vorschlag: Die Abweichungen vom Durchschnitt sind miteinander malzunehmen.
Sind beide hoch negativ oder hoch positiv, ergibt sich ein großes positives Gewicht.
Ist einer oder sind beide eher klein, ergibt sich ein kleines positives Gewicht.
Ist eine oder sind beide Null, ergibt sich ein Gewicht von Null.
Ist eine positiv und die andere negativ, so ergibt sich ein negatives Gewicht.
Diese Gewicht werden dann aufsummiert, durch die Anzahl der Wertepaare und die beiden Standardabweichungen geteilt, und voila: Da ist der berühmte Korrelationskoeffizient.“

Formel 1 für den Korrelationskoeffizienten
Der Text von Krämer kann unmittelbar in eine passende Formel gebracht werden:

$$r = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Für einen einfachen Fall wird der Korrelationskoeffizient mit dieser Formel handschriftlich berechnet:



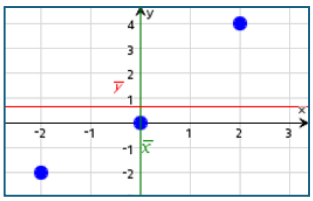
$\bar{x} = \frac{-1+0+1}{3} = 0\ \bar{y} = \frac{-1+0+2}{3} = \frac{1}{3}$

$\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{3} \cdot \left((-1) \cdot (-1) + 0 \cdot 0 + 1 \cdot 2\right) = 1$

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

$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\frac{1}{3} \cdot \left[\left(-\frac{4}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(\frac{5}{3}\right)^2\right]} = \sqrt{\frac{14}{9}} \Rightarrow \boldsymbol{r = \frac{1}{\sqrt{\frac{2}{3}} \cdot \sqrt{\frac{14}{9}}} \approx 0,98}$

Es soll nun exemplarisch klargemacht werden, weshalb für die Definition des Korrelationskoeffizienten die Division des Produkts der Abweichungen der x- und y-Werte von ihren Mittelwerten durch das Produkt der Standardabweichungen der x- und der y-Werte sinnvoll ist:



Alle x-Werte und alle y-Werte der Punkte vom ersten Beispiel wurden verdoppelt. Ihre gegenseitige Lage ist nach der Verdopplung geometrisch ähnlich zur vorherigen Situation.

$\bar{x} = \frac{-2+0+2}{3} = 0\ \bar{y} = \frac{-2+0+4}{3} = \frac{2}{3}\ \ \frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{3} \cdot \left((-2) \cdot \left(-\frac{8}{3}\right) + 0 \cdot \frac{2}{3} + 2 \cdot \frac{10}{3}\right) = 4$

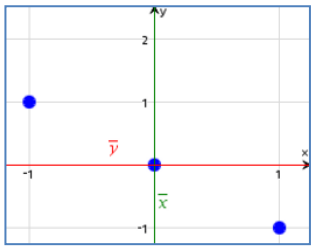
$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{3} \cdot [(-2)^2 + 0^2 + 2^2]} = \sqrt{\frac{8}{3}} = 2 \cdot \sqrt{\frac{2}{3}}$

$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\frac{1}{3} \cdot \left[\left(-\frac{8}{3}\right)^2 + \left(-\frac{2}{3}\right)^2 + \left(\frac{10}{3}\right)^2\right]} = \sqrt{\frac{56}{9}} = 2 \cdot \sqrt{\frac{14}{9}} \Rightarrow$

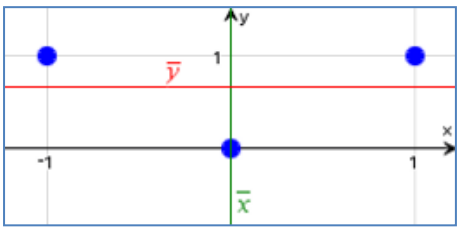
$\boldsymbol{r = \frac{4}{2 \cdot \sqrt{\frac{2}{3}} \cdot 2 \cdot \sqrt{\frac{14}{9}}} = \frac{1}{\sqrt{\frac{2}{3}} \cdot \sqrt{\frac{14}{9}}} \approx 0,98}$

Ohne die Division durch das Produkt der Standardabweichungen würde sich das Produkt der Abweichungen der x- und y-Werte von ihren Mittelwerten durch die Vergrößerung der Abstände zwischen den Punkten verändern. Die Division durch das Produkt der Standardabweichungen der x- und der y-Werte erzwingt eine Normierung. Der Korrelationskoeffizient r bleibt dadurch gegenüber der ersten Situation (Seite 3 oben) unverändert.

Für die beiden nächsten Fälle werden nur die Ergebnisse angegeben. Interessierte Leserinnen und Leser können das gern handschriftlich nachrechnen.



$$r = \frac{0}{\sqrt{\frac{2}{3}} \cdot \sqrt{\frac{2}{9}}} = 0$$



$$r = \frac{0}{\sqrt{\frac{2}{3}} \cdot \sqrt{\frac{2}{9}}} = 0$$

Wie sich schon an diesen einfachen Beispielen sehen lässt, ist die Anwendung der Formel 1 ziemlich aufwendig. Im Anhang finden Sie eine Umformung der Formel 1 für den Korrelationskoeffizienten r in eine Gestalt, die für die Anwendung durch den TI-30X Prio MathPrint™ hilfreich ist, weil sich damit die Rechnung mit dem Taschenrechner einfacher bewältigen lässt. Diese Formel wird nun angegeben und verwendet.

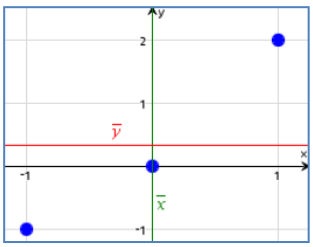
Formel 2 für den Korrelationskoeffizienten (Beweis im Anhang):

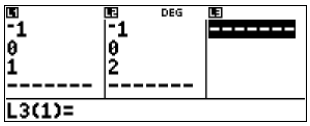
$$r = \frac{\sum xy - \frac{\sum x \cdot \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \cdot \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

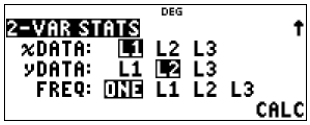
Für die Rechnung mit dem TI-30X Prio MathPrint™ sind dabei folgende Hinweise hilfreich:

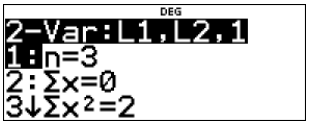
- Unter [data] werden die Werte für die x-Koordinaten (L1) bzw. y-Koordinaten (L2) in den Listeneditor übernommen.
- Mit [2nd][data][3] wird die Zwei-Variablenstatistik für L1 und L2 aufgerufen.
- Den angegebenen Kenngrößen lassen sich die zur Anwendung der Formel notwendigen Werte entnehmen.
- Um eine Kenngröße in den Hauptbildschirm zu übertragen, wird die Taste mit der Zahl vor dem Term gedrückt. Dabei wird „im Hintergrund“ auch der zugehörige Wert aktiviert.
  Beispiel: Um $\sum xy$ in den Hauptbildschirm zu kopieren, wird 6 gedrückt.
- Mit [2nd][data][1] geht es zurück in die 2-Variablenstatistik, um die nächste benötigte Kenngröße zu übertragen.
- Hinweis: Sollte man mit [2nd][data][1] nicht in den Bildschirm mit den Kenngrößen der Zwei-Variablenstatistik gelangen, kann dieser mit [2nd][data][3] wieder aufgerufen werden.

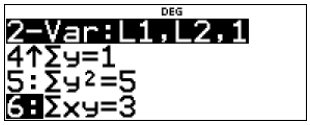
Im Folgenden wird die Anwendung auf das Beispiel von Seite 3 oben gezeigt. Dabei werden einige Teilterme des Ausdrucks einzeln berechnet und deren Ergebnisse unter Variablen gespeichert. Das Endergebnis wird dann mithilfe dieser Variablen ermittelt-

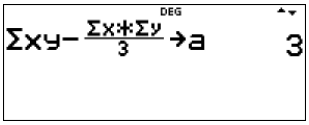


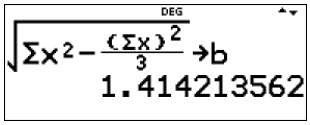


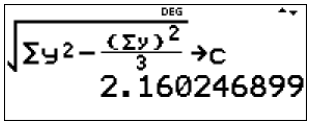


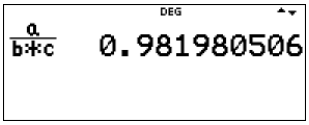








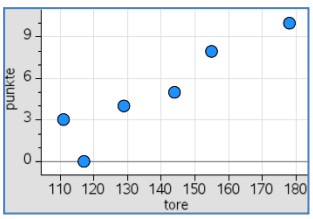


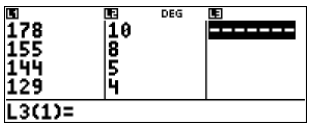


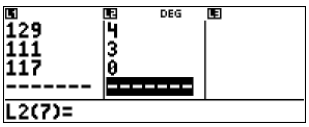
Der Korrelationskoeffizient hat wie auf Seite 3 den Wert $r \approx 0{,}98$.
Dies entspricht einem hoch positiven Zusammenhang.

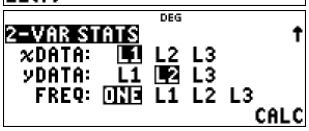
Analog erhält man auf diesem Wege für die beiden anderen Beispiele von Seite 2 und 3 ebenfalls die dort ermittelten Werte für r.

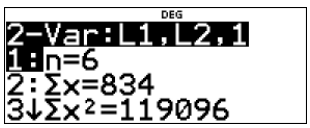
Auf etwas kürzerem Wege wird der Korrelationskoeffizient für das Eingangsbeispiel von Seite 1 für den Zusammenhang zwischen den erzielten Toren und den erreichten Punkten berechnet. Der Term für Formel 2 wird als Ganzes nach und nach in den Hauptbildschirm kopiert und ausgewertet:

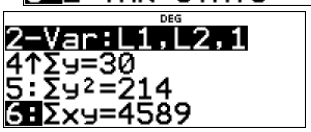


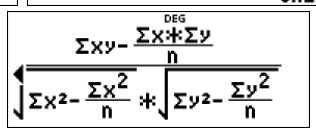


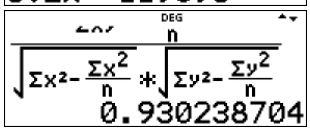








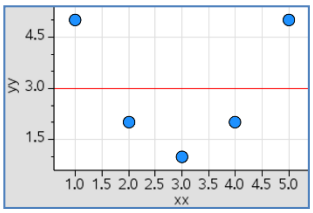




Wegen $r \approx 0{,}93$ liegt ein stark positiver Zusammenhang zwischen der Anzahl der erzielten Tote und der Anzahl der erreichten Punkte vor.

Bemerkungen zur Interpretation des Korrelationskoeffizienten r:
Es gilt stets $-1 \leq r \leq 1$.
Gilt r = + 1, so besteht ein perfekter positiver, für r = - 1 ein perfekter negativer linearer Zusammenhang zwischen den beiden Merkmalen.
Liegt r nahe bei + 1 bzw. nahe bei – 1, so spricht man von einem hohen positiven bzw. hohen negativen Zusammenhang beider untersuchter Merkmale.
Ein Wert von z. B. + 0,6 bedeutet, dass ein mittlerer positiver Zusammenhang besteht, ein Wert von z. B. - 0,2, dass ein kleiner negativer Zusammenhang vorliegt.
Gilt r = 0, so besteht kein linearer Zusammenhang zwischen den beiden Merkmalen.

Es ist aber durchaus möglich, dass ein Korrelationskoeffizient r = 0 zwischen zwei Variablen zwar nicht mit einem linearen Zusammenhang einhergeht, aber es durchaus einen nichtlinearen Zusammenhang zwischen ihnen geben kann. Das nebenstehende Diagramm veranschaulicht einen solchen Fall.



Hierbei ist zu erwähnen, dass der Korrelationskoeffizient zwar etwas über die Korrelation aussagt, sich aus dem Ergebnis aber nicht zwingend ein kausaler Zusammenhang ableiten lässt.

Ein bekanntes Beispiel für eine solche „Scheinkorrelation“ ist die Korrelation zwischen der menschlichen Geburtenrate und der Zahl der Storchenpaare in verschiedenen europäischen Regionen (s. Wikipedia). Obwohl es eine Korrelation zwischen der Zahl der Geburten und der Zahl der Storchenpaare gibt (d. h. mehr Geburten und gleichzeitig mehr Storchenpaare), gibt es keinen kausalen Zusammenhang (die falsche Schlussfolgerung, dass die Kinder vom Storch gebracht werden). Die Korrelation zwischen Geburten und Storchpaaren ergibt sich daraus, dass in ländlichen Regionen mehr Störche nisten und tendenziell auch mehr Kinder pro Paar geboren werden.

Anhang

Die Formel 1 kann vereinfacht werden. Zur Vereinfachung der Schreibweise werden im Folgenden die Indizes für die Zählvariable beim Summenzeichen weggelassen.

Zähler:

$$\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})=\frac{1}{n}\cdot\sum(x_i\cdot y_i-\bar{y}\cdot x_i-\bar{x}\cdot y_i+\bar{x}\cdot\bar{y})=$$

$$=\frac{1}{n}\cdot\sum x_i\cdot y_i-\bar{y}\cdot\frac{1}{n}\cdot\sum x_i-\bar{x}\cdot\frac{1}{n}\cdot\sum y_i+\frac{1}{n}\cdot\sum\bar{x}\cdot\bar{y}$$

$$=\frac{1}{n}\cdot\sum x_i\cdot y_i-\bar{y}\cdot\bar{x}-\bar{y}\cdot\bar{x}+\frac{1}{n}\cdot n\cdot\bar{y}\cdot\bar{x}=\frac{1}{n}\cdot\sum x_i\cdot y_i-\bar{y}\cdot\bar{x}$$

Nenner:

Mit $\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\cdot\sum x_i^2-2\cdot\frac{1}{n}\bar{x}\cdot\sum x_i+\frac{1}{n}\cdot\sum\bar{x}^2=\frac{1}{n}\cdot\sum x_i^2-2\cdot\bar{x}^2+\bar{x}^2=\frac{1}{n}\cdot\sum x_i^2-\bar{x}^2$

und entsprechend für den zweiten Faktor folgt:

$$\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}=\sqrt{\frac{1}{n}\cdot\sum x_i^2-\bar{x}^2}\cdot\sqrt{\frac{1}{n}\cdot\sum y_i^2-\bar{y}^2}$$

Für den Korrelationskoeffizienten ergibt sich:

$$r=\frac{\frac{1}{n}\cdot\sum x_i\cdot y_i-\bar{y}\cdot\bar{x}}{\sqrt{\frac{1}{n}\sum x_i^2-\bar{x}^2}\cdot\sqrt{\frac{1}{n}\sum y_i^2-\bar{y}^2}}=\frac{E(X\cdot Y)-E(X)\cdot E(Y)}{\sqrt{E(X^2)-E(X)^2}\cdot\sqrt{E(y^2)-E(y)^2}}=\frac{\frac{\sum xy}{n}-\frac{\sum x}{n}\frac{\sum y}{n}}{\sqrt{\frac{\sum x^2}{n}-\left(\frac{\sum x}{n}\right)^2}\cdot\sqrt{\frac{\sum y^2}{n}-\left(\frac{\sum y}{n}\right)^2}}$$

$$r=\frac{\frac{1}{n}\left(\sum x\cdot y-\frac{\sum x\cdot\sum y}{n}\right)}{\sqrt{\frac{1}{n}\left(\sum x^2-\frac{(\sum x)^2}{n}\right)}\cdot\sqrt{\frac{1}{n}\left(\sum y^2-\frac{(\sum y)^2}{n}\right)}}=\frac{\sum x\cdot y-\frac{\sum x\cdot\sum y}{n}}{\sqrt{\sum x^2-\frac{(\sum x)^2}{n}}\cdot\sqrt{\sum y^2-\frac{(\sum y)^2}{n}}}$$

Literaturverzeichnis

Krämer, Walter: in „Statistik für alle“, Springer Heidelberg 2015
Wikipedia: http://de.wikipedia.org/wiki/Scheinkorrelation (zuletzt eingesehen am 24.03.2025)

Autor:
Dr. Wilfried Zappe