Die Korrelation von Merkmalen

In der Analyse von Datenmaterial ist eines der Hauptziele der Statistik eine Abhängigkeit bzw. einen Zusammenhang zwischen Merkmalen zu erkennen.

Die Korrelation ermittelt den **Grad der Stärke** der Abhängigkeit zwischen zwei Merkmalen.

Ein Maß für die lineare Unabhängigkeit zweier Variablen

Um den linearen Zusammenhang zweier Variablen X und Y zu bestimmen ist die Kovarianz nicht aussagekräftig genug, da ihr absoluter Wert abhängig von der Skalierung der Variablen ist.

Die *Korrelation* ist ein normiertes Maß für den linearen Zusammenhang zweier Variablen. Es gilt

$$r_{XY} = \frac{Cov(X,Y)}{s_X s_Y} = \frac{\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}}{s_X s_Y},$$

d.h. r_{XY} ist die durch das Produkt der Varianzen von X und Y normierte Kovarianz von X und Y.

r_{XY} ist auch anders anschreibbar:

$$\begin{aligned} r_{XY} &= +/- \sqrt{\left(\frac{Cov(X.Y)}{s_X s_Y}\right)^2} \\ \Rightarrow |r_{XY}| &= +\sqrt{\frac{Cov(X,Y)}{s_X^2} * \frac{Cov(X,Y)}{s_Y^2}} = \sqrt{b_{YX} b_{XY}} \end{aligned}$$

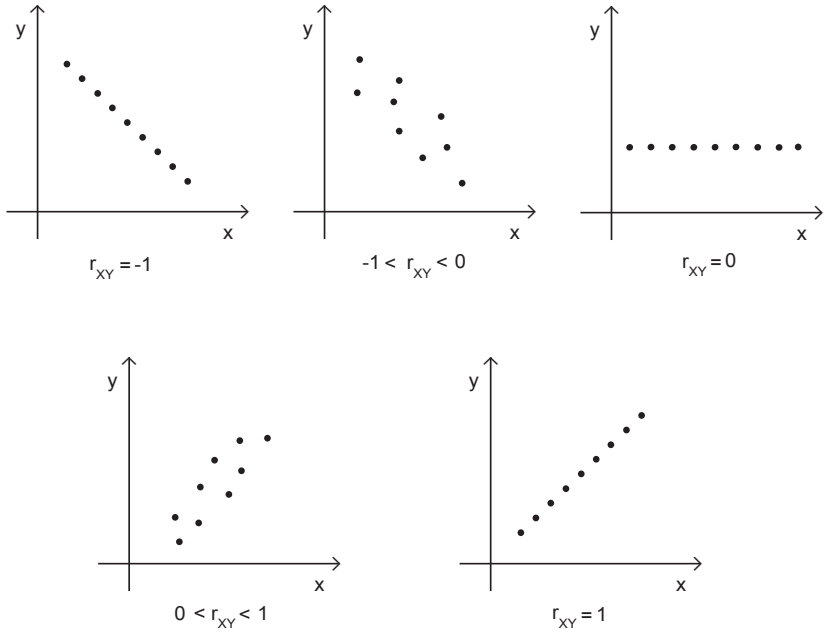
Das Vorzeichen von r_{XY} ist gleich dem Vorzeichen von $Cov(X,Y)$.

r_{XY} wird auch *Produktmomentkorrelation* genannt.

Beispiel: Verschiedene Datensätze der Grösse $n = 9$ werden im folgenden Plot grafisch dargestellt. Man sieht Beispiele dafür, wie Datensätze für die fünf Fälle

- $r_{XY} = -1$
- $-1 < r_{XY} < 0$
- $r_{XY} = 0$
- $0 < r_{XY} < 1$
- $r_{XY} = 1$

im Prinzip aussehen.



Eigenschaften der Korrelation

- Die Größe der Korrelation:
  - Die Korrelation r_{XY} ist 1 oder -1, wenn X und Y exakt linear abhängig sind, d.h.

    $$Y = cX + d, \quad c \neq 0,$$

    bzw.

    $$X = \frac{1}{c}Y - \frac{d}{c},$$

    d.h. die Regressionskoeffizienten b_{YX} bzw. b_{XY} sind c bzw. $1/c$, somit ist

    $$r_{XY} = +/- \sqrt{c * \frac{1}{c}} = +/- 1$$

    Auch hier ist das Vorzeichen von r_{XY} abhängig von der Kovarianz von X und Y
  - Wenn die Variablen linear unabhängig sind, so ist die Korrelation als auch die Kovarianz gleich 0.

Je höher der Betrag von r_{XY}, je höher ist die lineare Abhängigkeit zwischen X und Y. Der maximale Betrag der Korrelation ist 1, im Fall der exakten linearen Abhängigkeit.

Das Bestimmtheitsmaß

Wir betrachten zwei Variablen, X und Y. Wenn man Y als abhängige Variable von X betrachtet, so ist s_Y^2 die Gesamtvarianz und $s_{\hat{Y}}^2$ die durch die lineare Regression von X auf Y erklärte Varianz (zur Erinnerung: $\hat{Y}$ erhalten wir durch lineare Regression).

Wir definieren das Bestimmtheitsmaß B:

$$B = \frac{s^2_{\hat{Y}}}{s^2_Y}.$$

Das Bestimmtheitsmaß ist ein Maß für die Güte der Anpassung, die eine Regression erzielt. Es gilt stets $0 \leq B \leq 1$.

- B=1: d.h. $s^2_{\hat{Y}} = s^2_Y$ bzw. $y_i = \hat{y}_i$, alle Residuen sind also gleich 0 (das Residuum ist ja $y_i - \hat{y}_i$; es sind 100% der Varianz erklärt.
- B=0: wenn $s^2_{\hat{Y}} = 0$; es sind 0% der Varianz erklärt

Es besteht zwischen dem Bestimmtheitsmaß und r_{XY} ein Zusammenhang, und zwar folgender (ohne Beweis):

$$B = r^2_{XY}$$

Man kann r^2_{XY} bzw. B als als den Anteil der Varianz von Y, der durch X erklärt wird, verstehen (bzw. umgekehrt, je nachdem welche die abhängige Variable ist).

Ein Beispiel: Falls $r^2_{XY} = 0.49$, dann erklärt die Variable X die Variable Y zu 49%.

Die Varianz einer Summe

Mit dem Wissen, welches wir über die Kovarianz bzw. Korrelation erlangt haben, wollen wir folgende Fragen beantworten:
Wie kann man die Varianz einer Summe von Variablen aus Statistiken der Summanden darstellen? Sei also

$$Z = X + Y.$$

(die Messwerte sind $z_i = x_i + y_i$).

Man kann zeigen, dass folgende Gleichung gilt:

$$s_Z^2 = s_{X+Y}^2 = s_X^2 + s_Y^2 + 2Cov(X,Y).$$

Falls X und Y unkorreliert sind, dann gilt

$$r_{XY} = 0 \quad \text{bzw.} \quad Cov(X,Y) = 0.$$

In diesem Fall ist

$$s_Z^2 = s_X^2 + s_Y^2.$$

Die Rangkorrelation zweier Markmale

Möchte man die Korrelation zweier Merkmale X und Y, von denen man nur Ranginformationen besitzt, schätzen, so verwendet man den **Spearmanschen Rangkorrelationskoeffizienten** .

Eine weitere Methode liefert **Kendall**, auf die wir an dieser Stelle nicht weiter eingehen werden.

Beispiel: Zwei Personen (A und B) bewerten $n = 10$ Filme bezüglich ihres Unterhaltungswertes. Die Frage ist inwieweit die zwei Personen bezüglich ihrer Bewertung übereinstimmen, d.h. ob sie einen ähnlichen Filmgeschmack haben?

Die Rohwerte werden mit x_i (Person A) und y_i (Person B) bezeichnet. Die Skala ist angelehnt an das Schulnotensystem, d.h. 1 bedeutet der Film wird als sehr gut, bei 5 als sehr schlecht empfunden. Die Beurteilungstabelle ist folgende:

Film	A x_i	B y_i	A a_i	B b_i	d_i
Herr der Ringe I	1	1	2.0	1.5	0.5
Tiger and Dragon	2	3	5.0	4.5	0.5
E.T.	4	4	8.5	7.0	1.5
Mulholland Drive	1	5	2.0	9.5	-7.5
Blade Runner	1	3	2.0	4.5	-2.5
Fluch der Karibik	3	1	7.0	1.5	5.5
Wag the Dog	2	4	5.0	7.0	-2.0
Open Hearts	2	5	5.0	9.5	-4.5
Titanic	5	2	10.0	3.0	7.0
Battle Royale	4	4	8.5	7.0	1.5

a_i und b_i sind die Rangzahlen von A bzw. B. d_i ist die Differenz von a_i und b_i. Falls eine Bewertung öfters vorkommt, stellt sich die Frage, welche Rangzahl man in diesem Fall zuordnet. Wenn etwa die Bewertung 1 der Person A dreimal vorkommt, so nennt man dies eine "Bindung vom Ausmaß 3". Man bildet den Mittelwert der Rangzahlen die vergeben worden wären falls sich die Bewertungen unterscheiden würden, nämlich $1, 2$ und 3. D.h. der Mittelwert der Rangzahlen ist in diesem Fall

$$\frac{1+2+3}{2} = 2.$$

Herr der Ringe I,Blade Runner und Mulholland Drive erhalten demnach jeweils $a_i = 2.0$.

Berechnung der Rangkorrelation nach Spearman: Der statistische Zusammenhang zwischen A und B wird durch die Produktmomentkorrelation der a_i und b_i ausgedrückt. Wir bezeichnen diese *Rangkorrelation nach Spearman* mit r'.

Es gilt (ohne Angabe eines Beweises)

$$r' = 1 - \frac{6 \sum_i d_i^2}{n^3 - n}$$

Anmerkung: Für die Praxis gilt dass man diese Formel nur dann verwenden sollte, falls die Anzahl der Bindungen und deren Ausmaße klein sind.

In unserem Beispiel ergibt sich

$$\begin{aligned}
\sum_i d_i^2 &= 0.5^2 + 0.5^2 + 1.5^2 + (-7.5)^2 + (-2.5)^2 + 5.5^2 + (-2)^2 + (-4.5)^2 + 7^2 + 1.5^2 \\
&= 171 \\
&\Downarrow \\
r' &= 1 - \frac{6 \sum_i d_i^2}{n^3 - n} = 1 - \frac{6 * 171}{10^3 - 10} = 1 - \frac{1026}{990} = -0.036
\end{aligned}$$

Den Wert -0.036 liegt relativ nahe bei 0, d.h. die Personen A und B haben nach Spearman einen stark unterschiedlichen Filmgeschmack.