7: Korrelation

Statistische Verfahren in der Geographie

Till Straube <straube@geo.uni-frankfurt.de>

Institut für Humangeographie
Goethe-Universität Frankfurt

1 Lernziele dieser Sitzung

Sie können…

- ein Streudiagramm interpretieren.
- die Kovarianz von zwei Variablen berechnen.
- den Korrelationskoeffizienten von zwei Variablen berechnen.

2 Bivariate Statistik

Grundlage der bivariaten Statistik ist es, dass für eine Reihe von Untersuchungsheinheiten jeweils zwei Merkmale erfasst sind.

Diese Merkmale werden üblicherweise mit x und y gekennzeichnet. Für jedes i (laufende Nummer der Merkmalsträger*innen) gibt es dann ein x_i (Ausprägung des Merkmals x) und ein y_i (Ausprägung des Merkmals y).

Das Streudiagramm (engl. *scatter plot*) stellt alle erfassten Werte dar, indem es die Untersuchungseinheiten als Punkte arrangiert – und zwar anhand ihres jeweiligen Werts der Variable x entlang der x-Achse und entlang der y-Achse anhand des y-Werts (s. Abbildung 1).

2.1 Beispiel

Die statistischen Verfahren dieser Sitzung sollen wieder an einem Beispiel illustriert werden.

Wir fragen uns, ob der jährliche Ertrag in einem bestimmten Anbaugebiet für Klebreis in Nordostthailand mit dem jährlichen Niederschlag zusammenhängt. Die erfassten Werte sind in Tabelle 1 festgehalten („Rai“ ist ein in Thailand übliches Flächenmaß).

In einem Streudiagramm können diese Werte veranschaulicht werden. Dabei ist es üblich, die unabhängige Variable auf der x-Achse und die abhängige Variable auf der y-Achse einzutragen. Im Beispiel liegt nahe, dass der Ertrag vom Regen abhängt, und nicht etwa umgekehrt.

Abbildung 1 ist das Streudiagramm für unser Beispiel. Es fällt schon rein optisch auf, dass ein Zusammenhang zu bestehen scheint: Je mehr Regen, desto reicher die Ernte. Doch wie lässt sich dieser Zusammenhang beziffern?

Tabelle 1: Niederschlag und Ertrag im Reisanbau

Laufende Nr.	Jahr	Niederschlag (mm)	Ertrag (kg/Rai)
i		x_i	y_i
1	2008	1449	1860
2	2009	1472	2118
3	2010	1607	2225
4	2011	1494	2172
5	2012	1390	1816
6	2013	1764	2430
7	2014	1767	2580
8	2015	1765	2563
9	2016	1671	2276
10	2017	1838	2455

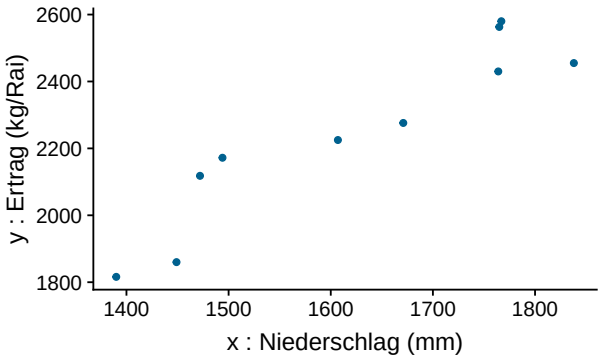

Abbildung 1: Streudiagramm zum Reisanbau

3 Kovarianz

Die Kovarianz (engl. *covariance*) s_{xy} gibt an, inwiefern die beiden Variablen x und y *gemeinsam variieren*. Die Kovarianz ergibt sich durch die Summe der jeweiligen Produkte der Differenzen zu den Mittelwerten $(x_i - \bar{x})$ und $(y_i - \bar{y})$, geteilt durch $(n - 1)$. Die Formel lautet also:

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

Gleichung 1 lässt erahnen: Wenn sowohl x als auch y in die gleiche Richtung vom jeweiligen Mittelwert abweichen (also beide Differenzen positiv oder beide Differenzen negativ), dann ist das Produkt positiv, sonst ist es negativ. Eine positive Kovarianz lässt also auf einen positiven Zusammenhang schließen (je größer x, desto größer auch y), eine negative Kovarianz auf einen negativen Zusammenhang (je größer x, desto *kleiner* y).

Softwarehinweis

Der Befehl `cov()` berechnet die Kovarianz einer bivariaten Verteilung in R.

Tabelle 2: Hilfstabelle für die Berechnung der Kovarianz

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	1449	1860	-172,7	-389,5	67266,65
2	1472	2118	-149,7	-131,5	19685,55
3	1607	2225	-14,7	-24,5	360,15
4	1494	2172	-127,7	-77,5	9896,75
5	1390	1816	-231,7	-433,5	100441,95
6	1764	2430	142,3	180,5	25685,15
7	1767	2580	145,3	330,5	48021,65
8	1765	2563	143,3	313,5	44924,55
9	1671	2276	49,3	26,5	1306,45
10	1838	2455	216,3	205,5	44449,65
Summe:	**16217**	**22495**			**362038,5**

3.1 Beispiel

Es macht Sinn, eine Tabelle anzulgen, in der Teilrechenschritte durchgeführt werden. Tabelle 2 veranschaulicht dies.

Als Zwischenschritt müssen die Mittelwerte $\bar{x}$ und $\bar{y}$ berechnet werden, wofür die Summen der ersten beiden Spalten herangezogen werden können:

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{16217}{10} = 1621{,}7\end{aligned}$$

$$\begin{aligned}\bar{y} &= \frac{\sum_{i=1}^{n} y_i}{n} \\ &= \frac{22495}{10} = 2249{,}5\end{aligned}$$

Schließlich ergibt Einsetzen der Produktsumme in Gleichung 1 die Kovarianz:

$$\begin{aligned}s_{xy} &= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} \\ &\approx \frac{362038{,}5}{9} = 40226{,}5\end{aligned}$$

Die Kovarianz ist also $s_{xy} = 40226{,}5$. Was sagt uns diese Zahl? Zunächst ist sie positiv, womit wir von einer positiven Korrelation (je mehr Regen, desto mehr Ertrag) ausgehen können. Sie ist auch „irgendwie“ ziemlich groß, was einen deutlichen Zusammenhang nahelegt. Aber die Kovarianz ist abhängig vom Maßstab – wäre der Ertrag nicht in Kilogramm pro Rai, sondern (wie in Deutschland

üblich) in Dezitonnen pro Hektar angegeben, dann wäre die Zahl deutlich kleiner (2514,156 um genau zu sein). Wie lässt sich die Stärke der Korrelation also unabhängig von den Maßeinheiten angeben?

4 Korrelationskoeffizient

Der Korrelationskoeffizent r (auch Produkt-Moment-Korrelation, Bravais-Pearson-Korrelation, Pearsons r, engl. *correlation coefficient*) standardisiert die Kovarianz s_{xy} anhand der Standardabweichungen s_x und s_y. Die Formel lautet:

$$r = \frac{s_{xy}}{s_x \cdot s_y} \tag{2}$$

Durch diese Standardisierung kann der Korrelationskoeffizient nur noch Werte zwischen $r = -1$ (perfekte negative Korrelation) und $r = 1$ (perfekte positive Korrelation) annehmen. Ein Korrelationskoeffizient nahe $r = 0$ bedeutet, dass es keinen Zusammenhang zwischen den Variablen x und y gibt (s. Abbildung 2).

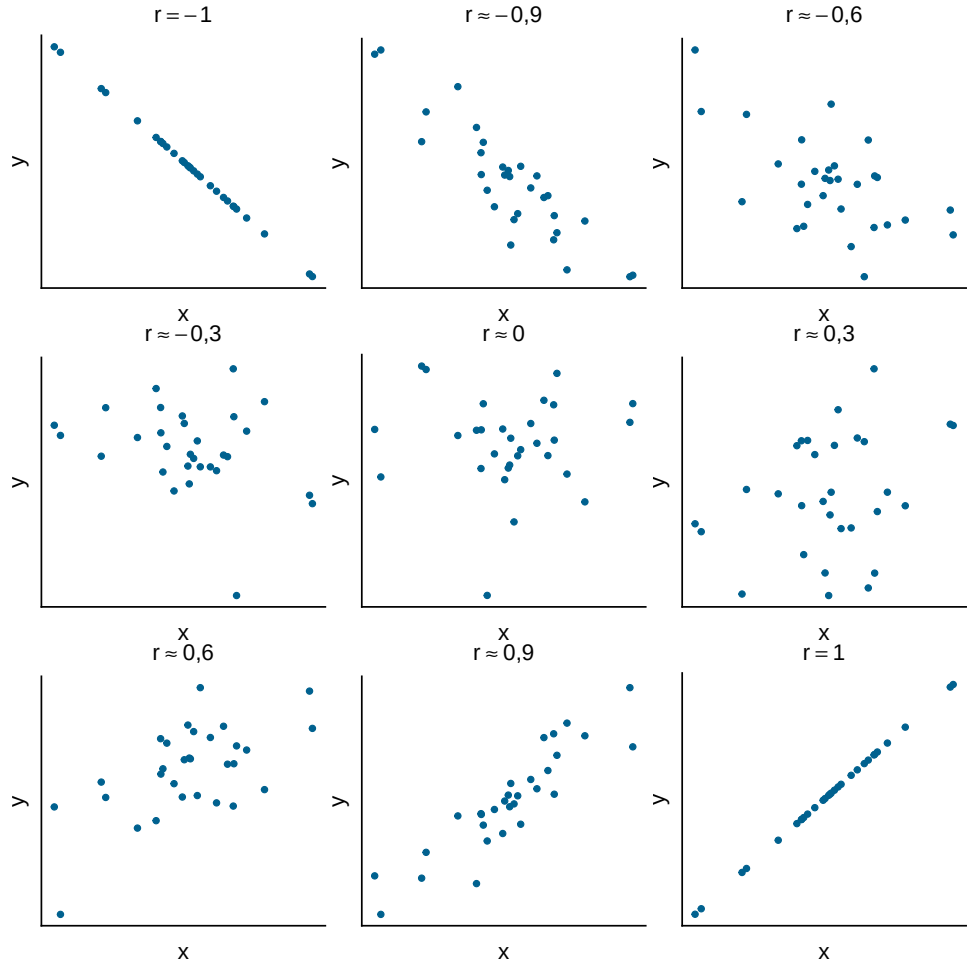

Abbildung 2: Verschiedene Korrelationskoeffizienten

Tabelle 3: Hilfstabelle für die Berechnung des Korrelationskoeffizienten

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	1449	1860	-172,7	-389,5	29825,29	151710,25
2	1472	2118	-149,7	-131,5	22410,09	17292,25
3	1607	2225	-14,7	-24,5	216,09	600,25
4	1494	2172	-127,7	-77,5	16307,29	6006,25
5	1390	1816	-231,7	-433,5	53684,89	187922,25
6	1764	2430	142,3	180,5	20249,29	32580,25
7	1767	2580	145,3	330,5	21112,09	109230,25
8	1765	2563	143,3	313,5	20534,89	98282,25
9	1671	2276	49,3	26,5	2430,49	702,25
10	1838	2455	216,3	205,5	46785,69	42230,25
Summe:	**16217**	**22495**			**233556,1**	**646556,5**

Softwarehinweis

In R kann der Korrelationskoeffizient von zwei Merkmalen mit dem Befehl `cor()` bestimmt werden.

4.1 Beispiel

In der Formel für den Korrelationskoeffizienten r (Gleichung 2) werden die Standardabweichungen s_x und s_y benötigt. Es ist daher sinnvoll, die Hilfstabelle um die Quadrate der Differenzen (und deren Summen) zu erweitern (s. Tabelle 3).

Die Standardabweichungen ergeben sich nun wie gewohnt aus:

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

$$= \sqrt{\frac{233556{,}1}{9}} = \sqrt{25950{,}68} \approx 161{,}09$$

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

$$= \sqrt{\frac{646556{,}5}{9}} = \sqrt{71839{,}61} \approx 268{,}03$$

Nun lassen sich die errechneten Werte in Gleichung 2 einsetzen:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$
$$\approx \frac{40226{,}5}{161{,}09 \cdot 268{,}03} \approx 0{,}93$$

Wir können bei einem Korrelationskoeffizienten $r \approx 0{,}93$ von einem deutlichen positiven Zusammenhang zwischen Niederschlag und Ertrag ausgehen.

5 Aufgaben

5.1 Aufgabe 1

Zeichnen Sie ein Streudiagramm und berechnen Sie die Kovarianz sowie den Korrelationskoeffizenten für die folgenden Messreihen.

a) Messreihe:

x_i	y_i
14,21	134
10,32	131
13,82	134
15,79	135
14,70	134
17,23	137
14,84	136
14,96	135

b) Messreihe:

x_i	y_i
-1,17	14,40
-0,10	2,31
-0,15	2,95
0,46	-1,39
0,34	-2,96
-0,44	2,44
2,13	-20,47
0,66	-10,51
-1,37	11,81
0,56	-4,05

5.2 Aufgabe 2

Sie erheben für zufällige „Wasserhäuschen“ in Frankfurt die Entfernung zur nächsten Haltestelle der S- oder U-Bahn sowie den durchschnitllichen Tagesumsatz. Die Erhebung ergibt:

Entfernung (m)	Umsatz (€/Tag)
35	394,61
79	468,92
234	385,75
105	376,17
318	283,26
31	342,77

Gibt es einen Zusammenhang zwischen Entfernung und Umsatz? Wenn ja: Wie hängen die Variablen zusammen? Wie stark ist der Zusammenhang?

5.3 Aufgabe 3

(weiterführend, nicht klausurrelevant… wirklich nur für Leute, die Spaß an Mathematik haben!)

a) Zeigen Sie, dass der Korrelationskoeffizent r ein standardisierter Wert ist, indem Sie ihn in z-Werten ausdrücken.
b) Überprüfen Sie die Formel anhand Aufgabe 1 a).
c) Angenommen, Sie wollen r angeben, ohne die Koviarianz berechnet zu haben. Wie lassen sich die Rechenschritte dann vereinfachen?
d) Überprüfen Sie den Rechenweg anhand Aufgabe 2.

6 Tipps zur Vertiefung

- YouTube-Kanal „Kurzes Tutorium Statistik“: Streudiagramm und Korrelation
- YouTube-Kanal „Methodenlehre Mainz“: Bivariate Daten (Playlist)
- Kapitel 10 in Bortz und Schuster (2010)
- Kapitel 6.1, 6.3 und 6.4 in Bahrenberg, Giese und Nipper (2010)
- Kapitel 16 in Klemm (2002)

Quellen

Bahrenberg, Gerhard, Ernst Giese und Josef Nipper. 2010. *Statistische Methoden in der Geographie*. 5. Aufl. Bd. 1. Univariate und bivariate Statistik. Stuttgart: Bornträger.

Bortz, Jürgen und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin: Springer.

Klemm, Elmar. 2002. *Einführung in die Statistik. Für die Sozialwissenschaften*. Wiesbaden: Westdeutscher Verlag.