STATISTIK I – Übung 08
Bravais-Pearson-Korrelationskoeffizient

1 Kurze Wiederholung

Was sind Zusammenhangsmaße?

Haben wir bislang immer nur eine Variable x (Lagemaße, Streuungsmaße, Schiefe und Wölbung, Grafiken etc.) individuell analysiert, sollen nun zwei Variablen x und y gleichzeitig betrachtet werden, um festzustellen, ob zwischen diesen ein Zusammenhang besteht. Dabei gilt, dass ein solcher Zusammenhang unterstellt werden kann – aber nicht existieren muss (siehe hierzu Aussagen zu Korrelation und Kausalität) – wenn die Variablen miteinander korrelieren (d.h. sich gleichmäßig zueinander verhalten). Hierbei wird in gleichsinnige Korrelationen (größere x-Werte fallen mit größeren y-Werten zusammen, kleinere x-Werte fallen mit kleineren y-Werten zusammen) und gegensinnige Korrelationen (größere x-Werte fallen mit kleineren y-Werten zusammen, kleinere x-Werte fallen mit größeren y-Werten zusammen) unterschieden. Die Stärke gleichsinniger oder auch gegensinniger Korrelationen (nicht aber die Stärke eines möglichen inhaltlichen Zusammenhangs) wird über sogenannte Zusammenhangsmaße wiedergegeben.

Welches Zusammenhangsmaß sich berechnen lässt, hängt – wie so vieles in der Statistik – vom Skalenniveau der betrachteten Variablen ab. Liegen nominalskalierte Variablen vor, eignet sich der Chi^2-Koeffizient, für ordinalskalierte Variablen empfiehlt sich dagegen der Rangkorrelationskoeffizient nach Spearman oder auch der Konkordanzkoeffizient nach Kendall, während sich für metrisch skalierte Daten wiederum der Bravais-Pearson-Korrelationskoeffizient anbietet. Wie in vielen Fällen ist die Berechnung der Parameter „abwärtskompatibel“, nicht aber „aufwärtskompatibel“, d.h. für metrisch skalierte Daten könnte beispielsweise auch der Chi^2-Koeffizient oder der Rangkorrelationskoeffizient nach Spearman berechnet werden, für ordinalskalierte Daten aber niemals der Bravais-Pearson-Korrelationskoeffizient. Grundsätzlich immer möglich ist auch die unterstützende grafische Zusammenhangsanalyse von Daten, wobei sich für diskrete Daten die Konstruktion gruppierter oder bedingter Balkendiagramme, für stetige Daten dagegen die Konstruktion von zwei- und dreidimensionalen Streudiagrammen sowie Scatterplot-Matrizen empfiehlt.

Korrelation und Kausalität

Von wesentlicher Bedeutung für die Interpretation von Zusammenhangsmaßen ist die Verinnerlichung der Tatsache, dass identifizierte Korrelationen zwar näher untersucht, niemals jedoch unmittelbar inhaltlich interpretiert werden sollten – auch dann nicht, wenn sich eine Interpretation im Hinblick auf einen sachlogischen Zusammenhang zwischen den betrachteten Variablen geradezu aufdrängt. Untersucht man beispielsweise den Zusammenhang zwischen Preis und Verkaufszahlen und identifiziert dabei eine starke Korrelation zwischen beiden Variablen, ist es zwar durchaus logisch davon auszugehen, dass der Preis die Verkaufszahlen beeinflusst (und nicht umgekehrt) – auf Basis eines Zusammenhangsmaßes lässt sich diese Annahme jedoch keinesfalls beweisen, da alle im Rahmen dieser Vorlesung betrachteten Zusammenhangsmaße keine Wirkungsrichtung (x beeinflusst y, y ist abhängig von x) kennen. Tatsächlich kann jede starke Korrelation zweier Variablen auf vier mögliche Ursachen zurückzuführen sein, die wir nachfolgend kurz betrachten wollen.

- Möglichkeit 1: Es liegt ein „echter“ inhaltlicher Zusammenhang mit einseitiger Wirkungsrichtung vor, d.h. x beeinflusst tatsächlich y bzw. y ist tatsächlich abhängig von x. Fälschlicherweise wird leider oft davon ausgegangen, dass eine starke Korrelation immer auf einen solchen monokausalen Zusammenhang zurückzuführen ist bzw. diesen sogar irgendwie „beweist“. Diesen Fehler sollte man (auch in der Klausur) nach Möglichkeit vermeiden.

- Möglichkeit 2: Es handelt sich um einen sogenannten beidseitigen Zusammenhang, d.h. x beeinflusst y während y gleichzeitig auch x beeinflusst. Dies ist etwa beim Bekanntheitsgrad eines Produktes und dessen Verkaufszahlen der Fall, da die Bekanntheit einerseits die Verkaufszahlen beeinflusst (ein stark beworbenes Produkt ist bekannter und wird häufiger angefragt), gleichzeitig aber auch die Verkaufszahlen den Bekanntheitsgrad beeinflussen (ein Produkt, das sich etwa über eine Niedrigpreisstrategie viel verkauft, wird mit der Zeit immer bekannter). Derartige beidseitige Zusammenhänge lassen sich nur mit statistischen Komplexverfahren näher untersuchen, die wir im Rahmen dieser Vorlesung nicht kennenlernen werden. Die besondere Schwierigkeit bei der Analyse beidseitiger Zusammenhänge besteht – wie man sich vermutlich leicht vorstellen kann – darin, festzustellen, welcher Anteil des Zusammenhangs sich auf welche Wirkungsrichtung (x -> y; y -> x) zurückführen lässt.

  Möglichkeit 3: Es handelt sich um einen rein zufälligen Effekt – eine sogenannte Scheinkorrelation. Korreliert man nur genügend Variablen gegeneinander, wird man immer Kombinationen finden, bei denen sich allein durch Zufall eine starke Korrelation ergibt. In der Vorlesung hatten wir hierzu die Internetseite des US-Juristen Tyler Vigen kennengelernt, der seit Jahren Korrelationen (ausgedrückt über den Bravais-Pearson-Korrelationskoeffizienten) sammelt, bei denen ein inhaltlicher Zusammenhang zwischen beiden Variablen mehr als unplausibel sein dürfte. Auf http://www.tylervigen.com/spurious-correlations finden sich viele hervorragende Beispiele dafür, warum eine – auch sehr starke – Korrelation noch lange nicht belegt, dass zwischen den betrachteten Variablen auch ein kausaler Zusammenhang existiert – so etwa zwischen der Scheidungsrate im US-Bundesstaat Maine und dem Pro-Kopf-Konsum von Margarine (r = 0,99258) oder dem Pro-Kopf-Konsum von Mozzarella und der Anzahl an Promovenden in den Ingenieurswissenschaften (r = 0,958648).

  Als Nebenbemerkung sei an dieser Stelle darauf hingewiesen, dass Untersuchungsdesigns, die auf dem wahllosen „Durchspielen“ einer Vielzahl von Korrelationsmöglichkeiten ohne die vorherige Aufstellung inhaltlich begründeter Vermutungen und Forschungsfragen basieren, aufgrund der Existenz derartiger Scheinkorrelationen abzulehnen sind. Testet man nämlich nur genügend Kombinationen von Variablen durch, werden sich letztlich immer starke oder sehr starke Korrelationen finden lassen, die jedoch nicht notwendigerweise auf einen inhaltlichen Zusammenhang zurückzuführen sind. Wir werden dieser Problematik bei der Betrachtung statistischer Testverfahren unter dem Stichwort „p-value hacking“ erneut begegnen.

- Möglichkeit 4: Es handelt sich um einen indirekten Zusammenhang zwischen x und y über eine dritte, nicht betrachtete Variable z. Das geradezu klassische Beispiel hierfür ist die über die Jahre mehrfach bestätigte, äußerst starke Korrelation zwischen der Geburtenrate und der Anzahl von Störchen in deutschen Landkreisen. Obwohl zwischen beiden Variablen – natürlich – kein unmittelbarer Zusammenhang besteht, sind sie doch über eine dritte Variable – den Grad der Urbanisierung – miteinander verbunden. Da der Urbanisierungsgrad sowohl die Geburtenquote (aus soziologischen Gründen) als auch die Storchdichte (aus ökologischen Gründen) gleichermaßen beeinflusst, korrelieren beide Variablen deutlich miteinander. Auch die Möglichkeit eines solchen Zusammenhangs sollte bei der Interpretation von Korrelationen nicht voreilig ausgeschlossen werden – bei betriebs- oder volkswirtschaftlichen Fragestellungen finden sich derartige „Hintergrundzusammenhänge“ sogar recht häufig.

Da eine starke Korrelation zweier Variablen theoretisch auf jede dieser vier Möglichkeiten zurückzuführen sein kann, sollte man sich vor der kausalen Interpretation von Korrelationen tunlichst hüten – auch wenn der inhaltliche Zusammenhang sich vermeintlich „offensichtlich" präsentiert. Eine Korrelation ist stets nur als ein Hinweis darauf zu begreifen, welche möglichen Zusammenhänge sich für eine vertiefte Analyse mit weiterführenden statistischen Verfahren (insbesondere Tests) anbieten. In vielen Vorlesungen und Lehrbüchern wird diese wirklich wesentliche Erkenntnis mit dem bekannten Merksatz „Korrelation ist nicht gleich Kausalität" vermittelt.

Der Bravais-Pearson-Korrelationskoeffizient

Liegen metrisch skalierte Daten (natürlich bei beiden Variablen) vor, kann – wie oben erwähnt – der Korrelationskoeffizient nach Bravais-Pearson berechnet werden. Dieser ist allerdings ausschließlich ein Maß für die Stärke einer linearen Korrelation zwischen zwei Variablen. Liegt eine andere Form des Zusammenhangs – beispielsweise ein quadratischer oder logarithmischer Zusammenhang – vor, wird dieser durch Bravais-Pearson nicht bzw. nicht vollumfänglich aufgedeckt. Bei der Interpretation des Korrelationskoeffizienten ist somit unbedingt zu berücksichtigen, dass ein niedriger Wert nicht bedeutet, dass zwischen den untersuchten Variablen keinerlei Zusammenhang besteht – er bedeutet lediglich, dass zwischen den untersuchten Variablen kein linearer Zusammenhang besteht.

Der Bravais-Pearson-Korrelationskoeffizient wird anhand folgender Formel berechnet:

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

Die Formel lässt erkennen, dass das Anlegen einer Hilfstabelle mit folgenden Spalten sinnvoll ist:

Nr.	x	y	x^2	y^2	(x * y)
1	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
n	...	...	...	...	...
Σ	...	...	...	...	...
Ø	...	...	//	//	//

Der Korrelationskoeffizient r liegt stets zwischen -1 und +1 und wird wie folgt interpretiert:

- Bei positiven Werten liegt ein positiver Zusammenhang vor (die Wertepaare liegen auf einer steigenden Geraden), bei negativen Werten ein negativer Zusammenhang (die Wertepaare liegen auf einer fallenden Geraden). Werte nahe Null deuten darauf hin, dass keine lineare Korrelation zwischen den beiden betrachteten Variablen vorliegt.

- Liegt der Betrag (!) von r nahe 0, liegt keine (lineare) Korrelation vor, bei Werten bis 0,5 kann von einer schwachen, bei Werten zwischen 0,5 und 0,8 von einer mittelstarken und bei Werten zwischen 0,8 bis 1,0 von einer starken Korrelation gesprochen werden, wobei der maximal erreichbare Betrag von 1,0 eine perfekte Korrelation (alle Wertepaare liegen aufgereiht auf einer Geraden) anzeigen würde. Die hier aufgezeigten Interpretationsgrenzen sind dabei nicht verbindlich, sondern stellen lediglich (erfahrungswertbasierte) Vorschläge dar.

2 Beispielrechnungen

Ein Eishändler variiert ceteris paribus (unter sonst gleichbleibenden Rahmenbedingungen) die Preise einer Kugel Eis über den Verlauf von sieben Tagen und zeichnet auf, wie viele Kugeln zu dem jeweiligen Preis veräußert wurden.

Tag	Kugelpreis (in Euro)	Verkaufte Kugeln (in Stück)
1	0,50	337
2	0,55	318
3	0,60	226
4	0,65	211
5	0,70	235
6	0,75	117
7	0,80	123

Berechnen Sie den Bravais-Pearson-Korrelationskoeffizienten.

Im ersten Schritt ist die bereits vorgestellte Hilfstabelle anzulegen und auszufüllen.

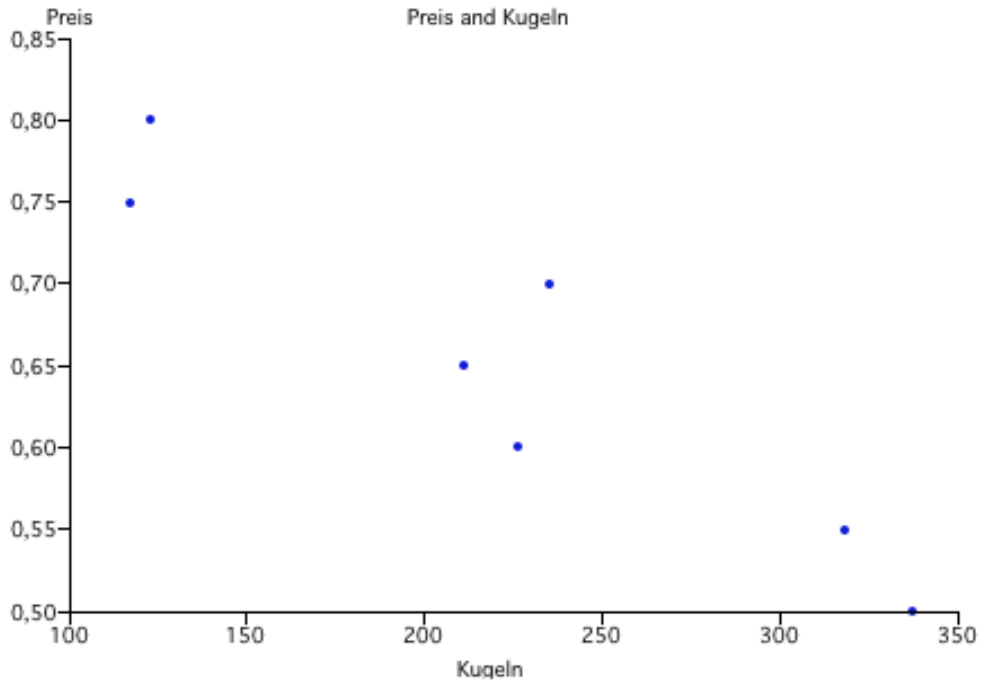
Nr.	x	y	x^2	y^2	(x * y)
1	0,50	337	0,25	113569	168,50
2	0,55	318	0,30	101124	174,90
3	0,60	226	0,36	51076	135,60
4	0,65	211	0,42	44521	137,15
5	0,70	235	0,49	55225	164,50
6	0,75	117	0,56	13689	87,75
7	0,80	123	0,64	15129	98,40
Σ	4,55	1567,00	3,03	394333	966,80
Ø	0,65	223,86	//	//	//

Die auf diese Weise errechneten Werte werden anschließend in die Formel eingetragen.

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

$$r = \frac{966{,}80 - 7 * 0{,}65 * 223{,}86}{\sqrt{3{,}03 - 7 * 0{,}65^2} * \sqrt{394333 - 7 * 223{,}86^2}} = -0{,}9372 \approx -0{,}94$$

Das Ergebnis weist auf eine starke negative Korrelation zwischen beiden Variablen hin – ein Ergebnis, das durch einen Blick auf das umseitige Streudiagramm (im Beispiel erstellt mit SSP) bestätigt wird.



3 Übungsaufgaben (Lösungen folgen in der kommenden Woche)

Eine Hochschule befragt zehn Absolventen/innen fünf Jahre nach ihrem Abschluss nach ihrem aktuellen Nettogehalt und ordnet die Summen den (metrisch skalierten!) Punkten in ihrer Abschlussklausur zu. Existiert eine lineare Korrelation zwischen dem Abschneiden in der Klausur und dem Nettogehalt nach fünf Jahren Berufstätigkeit?

Student Nr.	Klausurbewertung	Nettogehalt (in EUR)
1	98	2340
2	72	3750
3	65	1890
4	71	2420
5	56	2830
6	93	3110
7	66	1970
8	82	2480
9	89	2860
10	97	3333

Berechnen und interpretieren Sie den Bravais-Pearson-Korrelationskoeffizienten.