Ein neuer Weg, den Korrelationskoeffizienten ohne Bezug auf Kreuzprodukte zu lehren (oder berechnen)

SCHUYLER W. HUCK, BIXIANG REN UND HONGWEI YANG, KNOXVILLE –
ÜBERSETZUNG: MANFRED BOROVCNIK, KLAGENFURT

***Zusammenfassung**: Viele Studenten haben Schwierigkeiten, die begriffliche Querverbindung zwischen zweidimensionalen Daten, wie sie in einem Streudiagramm dargestellt werden, und der statistischen Kennziffer für den Zusammenhang, den Korrelationskoeffizienten* r*, zu sehen. Dieser Aufsatz zeigt, wie man* r *einführen (und berechnen) kann, indem der ‚direkte' und ‚indirekte' Einfluss jedes Punktes sichtbar gemacht wird; das führt zu einer neuen Formel für die Berechnung des Korrelationskoeffizienten* r.

Einleitung

In einem kürzlich in *Teaching Statistics* erschienenen Artikel mit dem Titel ‚Korrelation: Vom Schaubild zur Formel' streicht Peter Holmes (2001) sorgfältig heraus, dass Streudiagramme bei der Einführung in die Korrelation nützlich sind. Die Studierenden lernen rasch, aus solchen Bildern heraus einen Zusammenhang zwischen den zwei Variablen *X* und *Y* als (a) direkt oder indirekt bzw. als (b) stark, mittelmäßig oder schwach zu beurteilen.

Solche Urteile werden üblicherweise in Form einer Mutmaßung einer Zahl im Intervall zwischen –1 bis +1 ausgedrückt. Positiver und negativer Teil dieses Intervalls repräsentieren direkte bzw. indirekte Beziehungen; dabei ziehen starke (schwache) Beziehungen Werte von *r* nach sich, die nahe an den Intervallenden (in der Mitte) liegen. Studierende können bald recht genaue Schätzungen von *r* aus den Streudiagrammen ablesen.

Holmes berichtet, dass der typische Student – obwohl er angemessene Vorhersagen über das Vorzeichen und die Stärke einer Korrelation machen kann – Schwierigkeiten hat zu begreifen, wie die Formel von *r* ihr Ziel erreicht, nämlich das ‚qualitative Verständnis' aus der Betrachtung von Streudiagrammen zu quantifizieren. Er versucht, das ‚fehlende Bindeglied' einzufügen, indem er zeigt, wie die Formel für *r* Schritt für Schritt aus dem Streudiagramm hergeleitet werden kann.

Auch Russel Hurlburt versucht (2002, 382–385), die Lücke zwischen Streudiagramm und *r* zu schließen. Zuerst überlagert er das Streudiagramm mit einem Tic-Tac-Toe-Gitter (aus 9 Quadraten). Dann argumentiert er, dass die Daten in den vier Eck-Quadraten den größten Einfluss haben sowohl auf das Vorzeichen von *r* als auch auf dessen Größe. Schließlich berechnet Hurlburt das Kreuzprodukt der *z*-Werte für jeden Datenpunkt und erklärt, dass der Korrelationskoeffizient gleich dem Mittel dieser $z_x \cdot z_y$ - Werte ist, wobei

$$z_x = \sum (x - \bar{x}) / \sigma_x \text{ und } z_y = \sum (y - \bar{y}) / \sigma_y$$

sind. Die Notation σ_x zeigt dabei, dass die Standardabweichung mit Nenner n (und nicht $n-1$) berechnet wird; gleiches gilt für σ_y.

Sowohl Holmes als auch Hurlburt versuchen, das Streudiagramm mit der Formel für *r* über die Kreuzprodukte der *z*-Werte zu verbinden. Das geht natürlich so. Dennoch behauptet der Autor dieses Aufsatzes, dass es einen besseren Weg dazu gibt. Man kann wirklich zeigen, dass die Formel für *r* den qualitativen Eindruck von der Beziehung, den man aus dem Streudiagramm gewinnt, quantifiziert. Der Vorteil seines alternativen Zugangs ist, dass er sich nicht auf *z*-Werte bezieht; stattdessen werden für jeden Datenpunkt getrennte ‚direkte' und ‚indirekte' Komponenten erzeugt. Diese Komponenten, so wird behauptet, decken sich viel eher mit dem intuitiven ‚Gefühl', das man bekommt, wenn man Streudiagramme genauer betrachtet.

Strategie zum Unterrichten und Berechnen

Das Verfahren, den genauen ‚direkten' und ‚indirekten' Einfluss jedes Datenpunkts darzustellen, ist unkompliziert, leicht nachzuvollziehen und wird dem Ziel voll gerecht, die Art und Stärke eines zu untersuchenden Zusammenhangs zusammenzufassen. Das Verfahren ist in vier Schritte gegliedert:

Original 'A New Way to Teach (or Compute) Pearson´s *r* without Reliance on Cross-Products' in Teaching Statistics 29(2007)1, 13–16

Erstens, rechne die Rohdaten für x und y in z-Werte um. Diese Transformation beeinflusst den Wert von r nicht, denn der Korrelationskoeffizient ist invariant gegenüber linearen Transformationen von x und / oder y. Diese Eigenschaft des Korrelationskoeffizienten kann man leicht einsehen, wenn man folgende Frage stellt: ‚Wenn wir Körpergröße und Gewicht miteinander korrelieren, sollte dann die Wahl der Skala (Meter oder Zentimeter, Fuß oder Inch für die Länge; Unze, Pfund, Gramm oder Kilogramm für das Gewicht) die Größe des Korrelationskoeffizienten beeinflussen?

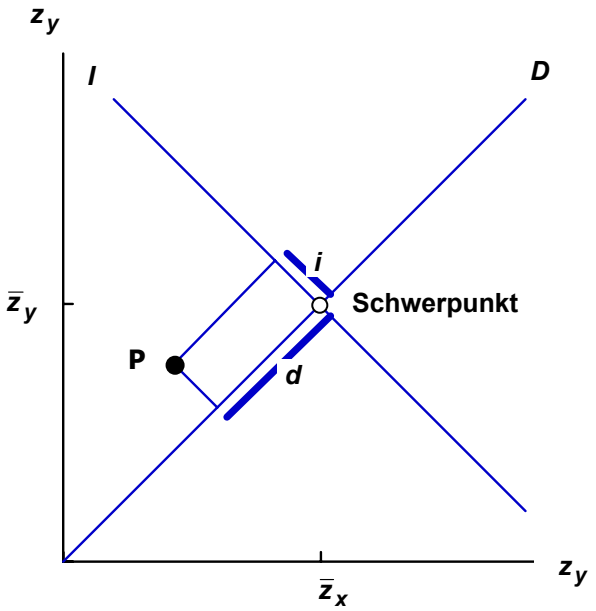

Abb. 1: Veranschaulichung der ‚direkten' (d) und ‚indirekten' (i) Beiträge eines Datenpunkts zu r

Zweitens, erstelle ein Streudiagramm mit den z-Werten aus Schritt 1 für die entsprechenden Punkte für die Daten. In dieses Streudiagramm zeichne man eine Gerade mit 45° Steigung, die durch den Schwerpunkt $(\bar{z}_x, \bar{z}_y) = (0;0)$ geht.

Bezeichne diese Gerade mit D (für ‚direkte Beziehung'. Zeichne dann eine zweite Gerade, senkrecht auf D durch den Schwerpunkt. Bezeichne diese zweite Gerade mit I (für ‚indirekte Beziehung').

Drittens, bestimme die Projektion jedes Datenpunkts auf die Geraden D und I; miss die Distanzen dieser Projektionspunkte vom Schwerpunkt, bezeichne diese Distanzen mit d und i. Der Wert von d zeigt den ‚direkten' Beitrag des Datenpunkts zu r an, während i den ‚indirekten' Beitrag zu r misst. Abb. 1 zeigt, wie dies für einen einzelnen Datenpunkt P getan wird.

Schließlich werden die Distanzen d und i, nachdem sie für alle Punkte berechnet sind, quadriert und summiert und in folgende Formel eingesetzt, um den Wert des Korrelationskoeffizienten r zu erhalten:

$$r = \frac{\sum d^2 - \sum i^2}{\sum d^2 + \sum i^2} \qquad (1)$$

Erläuterungen zur neuen Formel

Wie die Formel (1) zeigt, wird r positiv sein, wenn die d-Distanzen groß und die i-Distanzen klein sind. Dies tritt ein, wenn die Datenpunkte im Streudiagramm längs eines Pfads gruppiert sind, der sich von links unten nach rechts oben erstreckt. Ein schmälerer Pfad bringt mit sich, dass die i-Distanzen noch kleiner sind gegenüber den d-Distanzen, was verursacht, dass r sich gegen +1,00 hinbewegt. Umgekehrt bewegt sich r gegen –1,00 hin, wenn der Pfad der z-Werte sich eng längs der I-Geraden gruppiert. Man beachte, dass diese Ergebnisse für r leicht aus der obigen Formel abgelesen werden können, weil diese Formel mit dem ‚qualitativen' Eindruck übereinstimmt, den man aus der näheren Betrachtung des Streudiagramms gewinnt.

Einige Studierende werden sich wahrscheinlich wundern, warum es notwendig ist, die Werte von d und i zu quadrieren. Daher sollte der Lehrer vorbereitet sein, folgende Frage zu beantworten: ‚Warum kann man nicht einfach die d's addieren, die i's addieren, und dann die Differenz dieser beiden durch ihre Summe dividieren?' Egal ob vor oder nach dieser Frage, der Lehrer sollte herausstreichen, dass $\sum d$ und $\sum i$ jeweils den Wert 0 ergeben, für jeden Datensatz, ungeachtet der Stärke der Beziehung zwischen x und y. Daher sind $\sum d$ und $\sum i$ genauso unbrauchbar, um die Korrelation zweier Variablen zu messen, wie dies $\sum (x - \bar{x})$ (das sind die nicht-quadrierten Abweichungen vom Mittelwert) ist, wenn man die Streuung im eindimensionalen Fall erfasst.

Die Werte von d^2 und i^2 sind leicht zu berechnen, auch wenn sie sich eigentlich auf die zwei neuen, aufeinander senkrecht stehenden Achsen innerhalb des Streudiagramms beziehen (und nicht auf die senkrechte und waagrechte Achse, die mit z_x bzw. z_y bezeichnet ist). Eigentlich sind diese Werte für jeden beliebigen Datenpunkt eine ganz einfache Funktion von z_x und z_y:

$$d^2 = \frac{(z_x + z_y)^2}{2} \quad \text{und} \quad i^2 = \frac{(z_x - z_y)^2}{2}$$

x	y	$x-\bar{x}$	$y-\bar{y}$	z_x	z_y	d^2	i^2
7	4	3	–2	1,5	–0,5	0,5	2
5	0	1	–6	0,5	–1,5	0,5	2
4	8	0	2	0	0,5	0,125	0,125
3	6	–1	0	–0,5	0	0,125	0,125
1	12	–3	6	–1,5	1,5	0	4,5

$\bar{x} = 4$ $\bar{y} = 6$ $\sum (x-\bar{x}) \cdot (y-\bar{y}) = -30$ $\sum z_x \cdot z_y = -3{,}75$ $\sum d^2 = 1{,}25$ $\sum i^2 = 8{,}75$

$\sigma_x = 2$ $\sigma_y = 4$

$$r = \frac{\sum z_x \cdot z_y}{n} = \frac{-3{,}75}{5} = -0{,}75$$

$$r = \frac{\sum (x-\bar{x}) \cdot (y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2} \sqrt{\sum (y-\bar{y})^2}} = \frac{-30}{\sqrt{(20)(80)}} = -0{,}75$$

$$r = \frac{\sum d^2 - \sum i^2}{\sum d^2 + \sum i^2} = \frac{1{,}25 - 8{,}75}{1{,}25 + 8{,}75} = -0{,}75$$

Tab. 1: Ein Beispiel mit einfachen Daten zeigt, dass alle Formeln denselben Wert für *r* liefern

Auch wenn es seltsam erscheinen mag, dass sich die Ausdrücke auf d^2 und i^2 beziehen und nicht auf *d* und *i*, gibt es gute Gründe, warum wir das so getan haben. Ganz einfach gesprochen, die Vorzeichen der *d*- und *i*-Werte folgen schwerfälligen ‚Regeln'. Um genauer zu sein, ist das Vorzeichen von *d* für einen Datenpunkt positiv, wenn die zugehörigen *z*-Werte einer der folgenden drei Bedingungen genügt:

- sowohl z_x als auch z_y sind positiv
- z_y ist positiv, z_x negativ und $z_y > |z_x|$
- z_x ist positiv, z_y negativ und $z_x > |z_y|$

Ist keine der drei Bedingungen erfüllt, dann ist das Vorzeichen von *d* negativ.

Ähnliche Bedingungen kann man aufstellen, um das Vorzeichen von *i* zu bestimmen. Obwohl man diese Regeln aufstellen und daraus das Vorzeichen von *d* und *i* bestimmen kann, ist es viel leichter, d^2 und i^2 zu berechnen. Sowieso sind es nur diese Werte (und nicht von *d* und *i*), die man in unserer Formel (1) summieren muss.

Beispiel

Tab. 1 enthält fünf Paare von x- und y- Werten. Die Tabelle enthält auch die entsprechenden Abweichungen ($x-\bar{x}$, $y-\bar{y}$), *z*-Werte und die Werte für d^2 und i^2. Mit den Summenwerten dieser Tabelle kann man den Korrelationskoeffizienten auf drei verschiedene Arten berechnen. Zuerst mit der traditionellen Formel mit den Kreuzprodukten der *z*-Werte. Dann mit der üblichen Formel mit den Kreuzprodukten der Abweichungen vom Mittelwert. Schließlich mit der neuen Formel (1), die in diesem Aufsatz vorgestellt wurde. Wie man sehen kann, liefern alle drei Formeln denselben Wert für den Korrelationskoeffizienten.

Abschließende Bemerkungen

Alle einführenden Lehrbücher zur Statistik enthalten einen Abschnitt über Korrelation, in welchem die Autoren eines oder mehrere Streudiagramme zeigen und dann eine Formel für den Korrelationskoeffizienten *r* angeben. Man schaue etwa auf den Seiten 412–421 in Abrami u. a. (2001), 60–74 in Aron u. a. (2005), 263–275 in Caldwell (2004), 164–170 in McCall (2001) oder 131–135 in Rosenthal (2001) nach.

Diese und andere Autoren machen ihre Sache wirklich gut; sie erläutern Streudiagramme und zeigen, wie man den Korrelationskoeffizienten berechnet. Dennoch wären solche Darstellungen für Anfänger viel leichter zu verstehen, wenn der Zusammenhang zwischen dem Streudiagramm und der Formel für *r* auf Basis einer ‚intuitiven Geometrie' hergestellt würde. Die verwendete Bindung der Erklärung an Kreuzprodukte von *z*-Werten oder von Abweichungswerten ist nämlich nicht einfach zu verstehen.

Der klare Vorteil des Verfahrens, das hier skizziert wurde, liegt darin, dass der ‚direkte' und ‚indirekte' Einfluss jedes Datenpunkts begrifflich erfasst, gesehen und berechnet werden kann.

Man sollte anmerken, dass unsere Formel geradeso intuitiv ist wie das Streudiagramm selbst. Die gesamte Variabilität eines beliebigen Streudiagramms mit *z*-Werten könnte gemessen werden durch

(1) Bestimmen der linearen Distanz jedes Datenpunkts vom Schwerpunkt,

(2) Quadrieren dieser Distanzen, und

(3) Aufsummieren dieser Quadrate.

Dieses Maß an totaler Variabilität ist mathematisch äquivalent (über den Satz von Pythagoras) zum Nenner unserer Formel, $\sum d^2 + \sum i^2$; man kann sich diesen Wert vorstellen als Ergebnis von zwei unabhängigen ‚Kräften' – eine positiv und die andere negativ – welche die *n* Datenpunkte vom Schwerpunkt wegziehen.

Andererseits kann man sich den Zähler vorstellen als ein Maß der ‚Netto-Kraft', welche im Streudiagramm verbleibt, nachdem die kleinere dieser zwei entgegengesetzten Kräfte einen ebensolchen Anteil an der größeren Kraft eliminiert.

Der Quotient von Netto-Kraft und Totaler Kraft, die auf die Punkte einwirkt, ist daher ein geeigneter Weg, begrifflich zu fassen, was unsere Formel leistet.

Zwei zusätzliche Eigenschaften unseres Verfahrens sind es Wert, genannt zu werden, wenngleich wir empfehlen, dass Lehrer sich zurückhalten, dies mit Studierenden, die erst mit Korrelation anfangen, zu besprechen.

Erstens sind die beiden Geraden *D* und *I* in Abb. 1 identisch mit den zwei orthogonalen Komponenten einer Hauptkomponenten-Analyse der *z*-Werte. (Die erste Komponente einer Hauptkomponenten-Analyse wird mit *D* oder mit *I* übereinstimmen, je nachdem, ob *r* positiv oder negativ ist.).

Zweitens ist unsere Methode zur Berechnung von *r* in Formel (1) ganz analog zur Variation der Intra-Klassen-Korrelation, die man mit ICC $(3,1)$ bezeichnet. Bei zweidimensionalen *z*-Werten besteht der einzige Unterschied zwischen *r* und der Formel für ICC $(3,1)$ darin, dass ersterer Summen von Quadraten benutzt, während letzterer *mittlere* Quadrate verwendet. Trotz dieses Unterschieds geben beide Formeln identische Ergebnisse, wenn sie auf gepaarte standardisierte Werte (die *z*-Werte) angewendet werden.

Literatur

Abrami, P. C., Cholmsky, P. und Gordon, R. (2001): *Statistical Analysis for the Social Sciences*. Boston, Mass.: Allyn & Bacon.

Aron, A., Aron, E. N. und Coups, E. J. (2005): *Statistics for the Behavioural and Social Sciences* (3. Aufl.). Upper Saddle River, NJ: Prentice Hall.

Caldwell, S. (2004): *Statistics Unplugged*. Belmont, CA: Wadsworth.

Holmes, P. (2001): Correlation: From Picture to Formula. *Teaching Statistics* 23(3), 67–70.

Hurlburt, R. T. (2002): *Comprehending Behavioral Statistics* (3. Aufl.). New York: Brooks/Cole.

McCall, R. B. (2001): *Fundamental Statistics for the Behavioral Sciences*. Belmont, CA: Wadsworth.

Rosenthal, J. A. (2001): *Statistics and Data Interpretation for the Helping Professions*. Belmont, CA: Brooks/Cole.

Anschrift der Verfasser

Schuyler W. Huck, Bixiang Ren
und Hongwei Yang
University of Tennessee, Knoxville, USA
shuck@utk.edu