Johannes Barth

Woher stammt die 6 im Korrelationskoeffizienten von Spearman?

1. Vorbemerkung

In Statistikvorlesungen, -kursen und -übungen der deskriptiven Statistik werden u. a. die Zusammenhangsmaße von Pearson

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

und Spearman

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2 - 1)}$$

besprochen. Ihre mathematischen Formeln sind vom "Aussehen" her sehr verschieden. Des Öfteren taucht deshalb die Frage auf, ob beide Formeln etwas gemeinsam haben - die Antwort lautet ja. Sodann stellt sich die Frage, woher die 6 im Zähler der Formel von Spearman kommt - die Antwort lautet, das kann man irgendwie aus der Formel von Pearson ableiten. Möchte sich jetzt der interessierte Statistiker in Lehrbüchern schlau machen, so stößt er in der überwiegenden Zahl der Fälle nur auf die Darstellung der reinen Formeln.

Dies hat den Autor bewogen, den Zusammenhang zwischen dem Korrelationskoeffizienten von Pearson und dem Rangkorrelationskoeffizienten von Spearman etwas näher zu erläutern.

Die Frage nach dem (linearen) Zusammenhang, der Korrelation zwischen zwei Variablen ist in der Statistik von zentraler Bedeutung. Viele Statistiker haben sich mit diesem Thema beschäftigt und damit Eingang in die Lehrbücher der deskriptiven Statistik gefunden, so etwa Bravais-Pearson, Fechner, Spearman und Kendall.

2. Abhängigkeit von Merkmalen

Zur Charakterisierung zweidimensionaler (bivariabler) Häufigkeitsverteilungen kann man die Daten in Tabellen bzw. Kontingenztafeln erfassen, wobei das erste interessierende Merkmal mit X und das zweite mit Y bezeichnet wird. Aus rein methodischer Sicht ist es nicht notwendig zu unterscheiden, ob die Variable X von Y beeinflusst wird oder umgekehrt Y von X. Die grafische Darstellung kann in dreidimensionalen Bildern oder in einem x-y-Koordinatensystem anhand einer Punktwolke erfolgen.

Die Aufgabe in der Korrelationsrechnung besteht nun darin, den stochastischen (linearen) Zusammenhang zwischen den zu untersuchenden Variablen X und Y zu analysieren und durch geeignete Kennziffern, so genannte Zusammenhangsmaße, quantitativ zu charakterisieren, also Zusammenhänge zwischen stetigen oder auch diskreten Zufallsvariablen zu messen, deren Realisationen quantitativ oder qualitativ sein können. Hierbei soll darauf hingewiesen werden, dass die Statistik weder die Kausalität, d. h. ob eine Variable X von einer Variablen Y kausal beeinflusst wird oder umgekehrt Y von X, "beweisen" kann, noch stellt sie die Frage nach dem Sinn eines Zusammenhangs ("Scheinkorrelation"). Sie kann eine vorliegende Theorie - oder auch die vorhandenen Zweifel daran - bestenfalls "bestärken".

Die Wahl eines geeigneten Korrelationskoeffizienten hängt u. a. davon ab, welche Skalierung den Daten zu Grunde liegt, d. h. welche Sachlogik sich hinter den numerischen Merkmalsausprägungen verbirgt.

Im Folgenden werden die drei wichtigsten Skalierungen kurz erläutert.

Eine *Nominalskala* liegt vor, wenn die Ausprägungen des untersuchten Merkmals durch die zugeordneten Zahlen lediglich unterschieden werden sollen, d. h. sie lassen sich nicht eindeutig in einer Rangfolge ordnen (z. B. Geschlecht, Religion, Nationalität). Bei *ordinal skalierten* Merkmalen können die Ausprägungen nicht nur unterschieden, sondern auch in eine Rangfolge gebracht werden (z. B. Schulnoten, Intelligenzquotienten). Bei *metrisch skalierten* Merkmalen kann außer der Bildung der Rangfolge noch bestimmt werden, in welchem Ausmaß (Intervall) sich je zwei verschiedene Merkmalsausprägungen unterscheiden (z. B. Gewicht, Körpergröße, Entfernung, Produktionsdauer).

Bei nominal skalierten Daten werden die Zusammenhangsmaße Kontingenzkoeffizienten, bei ordinal skalierten Daten Rangkorrelationskoeffizienten und bei metrisch (kardinal) skalierten Daten (Maß-) Korrelationskoeffizienten genannt.

Für metrisch skalierte Daten liefert vor allem der von dem Statistiker K. Pearson entwickelte Korrelationskoeffizient r die gesuchte Information über den "Grad der Anschmiegung" der Beobachtungen an die Regressionsgerade oder, wie man auch sagt, über die "Strammheit des Zusammenhangs" zwischen den Beobachtungen der Variablen X und Y. Der Variationsbereich von r liegt zwischen - 1 und + 1.

Für den Fall aber, dass ordinale Datenreihen vorliegen, bzw. nur die Rangfolge von Daten angegeben werden kann, versagt die Berechnung des Abhängigkeitsmaßes nach Pearson. Um nun einen Zusammenhang bei solchen Datenreihen empirisch festzustellen, wurde von Spearman die Formel eines Rangkorrelationskoeffizienten r_s entwickelt. Er ist also ein ordinales Abhängigkeitsmaß und ebenfalls im Intervall $|r_s| \leq 1$ zu interpretieren. Der Rangkorrelationskoeffizient kann auch als Approximation für r eingesetzt werden. Er wird dann verwandt, wenn man den mit der Berechnung von r verbundenen Rechenaufwand scheut. Es muss aber immer berücksichtigt werden, dass r_s nur ein relativ grobes (Ersatz-) Maß für r darstellt und man auf die "Genauigkeit" des quantitativen Merkmals verzichtet. Es wird nur noch die Frage betrachtet, ob die den Merkmalsträgern zugrunde liegenden Werte größer oder kleiner sind. Da es bei vielen Problemstellungen nur auf Größenordnungen ankommt, kann r_s tatsächlich oft gute Dienste leisten.

Bei der Berechnung des Rangkorrelationskoeffizienten geht man so vor, dass man für die Ausgangsdaten Rangziffern vergibt, um nur noch mit diesen Rangziffern zu arbeiten, d. h. Originaldaten werden in Rangplätze umgewandelt. Bei der Festlegung der Rangzahlen ist darauf zu achten, dass gleich große Werte bei einer Veränderlichen (sog. Bindungen) auch mit gleichen Rangzahlen belegt werden. Um dies zu erreichen wird bei gleich großen Werten aus den Rangzahlen das arithmetische Mittel gebildet.

Eine statistisch exakte Beurteilung der Stärke des Zusammenhangs kann in Verbindung mit der Wahrscheinlichkeit erfolgen. Dann können auch Fragen nach einer "Signifikanz" des Zusammenhangs beantwortet werden.

3. Ableitung des Spearman-Korrelationskoeffizienten aus dem Korrelationskoeffizienten von Pearson

Für die Ableitung des Korrelationskoeffizienten von Spearman geht der Autor vom Korrelationskoeffizienten von Pearson in der mit (4) gekennzeichneten Form aus. Zur Illustration seien einige gleiche Formeln zur Berechnung des Pearson- Korrelationskoeffizienten nebeneinander gestellt.

Da alle Häufigkeiten gleich 1 sind, gelten folgende Formeln:

$$(1)\quad r = \frac{\operatorname{cov}(X,Y)}{s_x s_y} \qquad \text{(als normierte Kovarianz)}$$

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

$$(3)\quad = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n}y_i^2 - n\bar{y}^2)}}$$

$$(4)\quad = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\bar{y}}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2)(\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \bar{y}^2)}} \qquad \text{(als Quotient von Abweichungssummen)}$$

In Formel (4) werden nun an Stelle der Messwerte x_i und y_i Rangplätze $R(x_i)$ und $R(y_i)$ eingeführt. Der Einfachheit halber werden aber die Bezeichnungen x_i und y_i beibehalten.

Stillschweigend wird vorausgesetzt, dass die für die Berechnung von r notwendigen Voraussetzungen gegeben sind: Stetigkeit der beiden Zufallsvariablen X und Y sowie ihrer Verteilungen; möglichst bivariate Normalverteilung in der Grundgesamtheit; metrische Skalierung von X und Y und damit Äquidistanz (Gleichabständigkeit) aufeinander folgender Werte. Gerade die zuletzt genannte Voraussetzung ist bei Rangfolgen meist nie gegeben. Auch ist die Problematik, wie gleiche Rangplätze zu berücksichtigen sind, für die Berechnung des Spearman-Korrelationskoeffizienten nicht hinreichend genau geklärt. Für den Fall, dass keine Rangbindungen vorkommen, d. h. keine gleich großen Ränge vergeben werden müssen, lässt sich folgende Berechnung durchführen:

Bei n Beobachtungspaaren durchlaufen die Rangdaten x_i und y_i daher alle Werte von 1 bis n und es gilt

$$(5)\quad \bar{x} = \bar{y} \text{ und } s_x^2 = s_y^2.$$

Weiter gilt:

$$(6)\quad \sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i = \frac{1}{2}n(n+1)$$

sowie

$$(7)\quad \sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}y_i^2 = \frac{1}{6}n(n+1)(2n+1) \text{ mit } \bar{x} = \frac{\sum_{i=1}^{n}x_i}{n}$$

Für den Nenner der obigen Formel (4) des Korrelationskoeffizienten folgt damit:

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

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

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

$$=\frac{1}{n}\cdot\frac{1}{6}n(n+1)(2n+1)-(\frac{1}{n}\cdot\frac{1}{2}n(n+1))^2$$

$$=\frac{1}{6}(n+1)(2n+1)-(\frac{1}{2}(n+1))^2=\frac{1}{6}(2n^2+3n+1)-\frac{1}{4}(n^2+2n+1)$$

$$=\frac{1}{12}(4n^2+6n+2)-\frac{1}{12}(3n^2+6n+3)$$

$$=\frac{1}{12}(4n^2-3n^2+6n-6n+2-3)$$

$$(8)\quad =\frac{1}{12}(n^2-1)$$

Für den Zähler der Formel (4) des Korrelationskoeffizienten, also die Kovarianz, gilt die Beziehung

$$\sum_{i=1}^{n}(x_i-y_i)^2=\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_iy_i+\sum_{i=1}^{n}y_i^2$$

bzw.

$$\sum_{i=1}^{n}x_iy_i=\frac{1}{2}\sum_{i=1}^{n}x_i^2+\frac{1}{2}\sum_{i=1}^{n}y_i^2-\frac{1}{2}\sum_{i=1}^{n}(x_i-y_i)^2$$

und man erhält

$$\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{xy}$$

$$=\frac{1}{n}(\frac{1}{2}\sum_{i=1}^{n}x_i^2+\frac{1}{2}\sum_{i=1}^{n}y_i^2-\frac{1}{2}\sum_{i=1}^{n}(x_i-y_i)^2)-\frac{\sum_{i=1}^{n}x_i}{n}\cdot\frac{\sum_{i=1}^{n}y_i}{n}$$

$$=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\frac{1}{2n}\sum_{i=1}^{n}(x_i-y_i)^2-(\frac{\sum_{i=1}^{n}x_i}{n})^2$$

$$=\frac{1}{n}\cdot\frac{1}{6}n(n+1)(2n+1)-\frac{1}{2n}\sum_{i=1}^{n}(x_i-y_i)^2-(\frac{n(n+1)}{2n})^2$$

$$=\frac{1}{6}(2n^2+2n+n+1)-\frac{1}{2n}\sum_{i=1}^{n}(x_i-y_i)^2-\frac{n^2+2n+1}{4}$$

$$=\frac{1}{12}(4n^2+6n+2-3n^2-6n-3)-\frac{1}{2n}\sum_{i=1}^{n}(x_i-y_i)^2$$

$$(9)\quad =\frac{1}{12}(n^2-1)-\frac{1}{2n}\sum_{i=1}^{n}(x_i-y_i)^2$$

Wenn nun die Ausdrücke für Zähler (9) und Nenner (8) in die Formel (4) des Korrelationskoeffizienten r eingesetzt werden, so erhält man:

$$r=\frac{\frac{1}{12}(n^2-1)-\frac{1}{2n}\sum_{i=1}^{n}(x_i-y_i)^2}{\frac{1}{12}(n^2-1)}=1-\frac{6\sum_{i=1}^{n}(x_i-y_i)^2}{n(n^2-1)}=r_s$$

Man sieht also, der Rangkorrelationskoeffizient von Spearman r_s ist ein auf die Rangzahlen angewandter Pearson-Korrelationskoeffizient. Er ist ein Maß für den Zusammenhang zwischen Merkmalen, die nach einer Rangskala geordnet werden können. Die Differenz der Rangziffern schreibt man häufig als $D_i = x_i - y_i$ und nennt sie Rangdifferenzen.

4. Anwendungen

1. Beispiel

Das erste Beispiel sei die Berechnung des Zusammenhangs zwischen Wahlbeteiligung und den Anteilen der beiden Parteien CDU und SPD an den gültigen Stimmen bei der Landtagswahl 1999 im Saarland. Eine alte Wahlweisheit lautet, je höher die Wahlbeteiligung ist, desto besser ist das Ergebnis der CDU und desto niedriger das der SPD. Es soll überprüft werden, ob sich dieser Zusammenhang bei der Landtagswahl 1999 für die Ergebnisse der saarländischen Gemeinden bestätigt.

Auf der Grundlage der Ergebnisse der Landtagswahl 1999 nach Gemeinden werden die Zusammenhangsmaße nach Pearson und nach einer Umrechnung in Ränge nach Spearman berechnet. (Originaldaten findet der Leser auch im Inter-

Gemeinde	Wahlbeteil.	SPD	CDU	Wahlbeteil.	SPD	CDU
	Anteil in %			Rang		
Saarbrücken, Landeshauptstadt	62,8	45,0	40,5	3,5	29,5	9
Friedrichsthal, Stadt	70,8	53,5	39,7	22,5	51	8
Großrosseln	73,1	53,9	38,0	34,5	52	4
Heusweiler	70,7	42,4	48,5	20,5	19	35,5
Kleinblittersdorf	72,7	42,3	47,8	30,5	18	28
Püttlingen, Stadt	73,7	41,5	50,5	37	14	40
Quierschied	73,6	41,6	52,1	36	15	45
Riegelsberg	72,6	42,1	48,0	28,5	16	31
Sulzbach/Saar, Stadt	64,2	49,7	39,3	5	45	6
Völklingen, Stadt	62,8	50,0	39,6	3,5	47	7
Beckingen	71,7	46,1	45,6	25	33	20
Losheim am See	70,7	40,3	51,0	20,5	11	42
Merzig, Kreisstadt	68,6	42,5	49,2	14	20	37,5
Mettlach	67,9	45,0	47,1	10,5	29,5	26
Perl	72,5	32,8	59,6	27	1	52
Wadern, Stadt	74,7	42,9	48,3	42	21	33,5
Weiskirchen	74,3	38,9	54,5	41	9	50
Eppelborn	76,0	41,0	53,0	44	13	46
Illingen	76,7	43,8	47,9	46	25	29,5
Merchweiler	70,3	46,7	45,8	19	34,5	21
Neunkirchen, Kreisstadt	61,0	53,0	36,8	1	50	2
Ottweiler, Stadt	69,5	49,9	36,9	17	46	3
Schiffweiler	70,8	52,7	38,5	22,5	49	5
Spiesen-Elversberg	65,0	48,4	41,1	7	40	11
Dillingen/Saar, Stadt	65,6	43,2	48,2	8	23	32
Lebach, Stadt	73,1	37,9	54,0	34,5	4	48
Nalbach	72,8	38,2	53,3	32	5,5	47
Rehlingen-Siersburg	73,8	48,8	43,1	38,5	42	15
Saarlouis, Kreisstadt	64,8	42,2	47,9	6	17	29,5
Saarwellingen	68,2	44,6	46,8	12	27	24
Schmelz	70,2	43,5	48,5	18	24	35,5
Schwalbach	73,0	48,0	44,1	33	39	16,5
Überherrn	77,2	46,7	45,0	47	34,5	19
Wadgassen	67,2	47,9	42,9	9	38	13
Wallerfangen	68,7	45,5	44,1	15	31	16,5
Bous	72,1	48,9	40,7	26	43	10
Ensdorf	69,4	49,0	42,2	16	44	12
Bexbach, Stadt	68,5	46,9	43,0	13	36	14
Blieskastel, Stadt	71,5	40,5	47,3	24	12	27
Gersheim	77,4	38,6	51,4	49	8	43
Homburg, Kreisstadt	61,2	38,2	49,2	2	5,5	37,5
Kirkel	72,6	50,1	35,8	28,5	48	1
Mandelbachtal	76,5	37,0	50,6	45	3	41
St. Ingbert, Stadt	67,9	39,4	47,0	10,5	10	25
Freisen	75,0	44,9	49,6	43	28	39
Marpingen	81,2	47,3	46,5	51	37	23
Namborn	74,1	43,9	48,3	40	26	33,5
Nohfelden	77,3	48,6	44,6	48	41	18
Nonnweiler	78,8	45,7	46,0	50	32	22
Oberthal	81,9	43,0	51,5	52	22	44
St. Wendel, Kreisstadt	72,7	38,3	54,1	30,5	7	49
Tholey	73,8	34,3	58,2	38,5	2	51

net unter www.statistik.saarland.de oder in der Einzelschrift Nr. 106/99 "Wahlen 1999").

Die Berechnung des Korrelationskoeffizienten von Pearson ergibt einen Zusammenhang von $r = + 0{,}395$ zwischen Wahlbeteiligung und Stimmenanteil der CDU und von $r = - 0{,}208$ zwischen Wahlbeteiligung und Stimmenanteil der SPD.

Die entsprechenden Rangkoeffizienten lauten $r_s = + 0{,}418$ und $r_s = - 0{,}231$. Damit bestätigen beide Koeffizienten obige Behauptung. Die Ergebnisse weichen aber auf Grund der zuvor beschriebenen unterschiedlichen Berechnungsansätze der Korrelationskoeffizienten von Pearson und Spearman etwas voneinander ab.

2. Beispiel

Ein typisches Beispiel aus den Lehrbüchern der deskriptiven Statistik zur Darstellung des Rangkorrelationskoeffizienten von Spearman ist die Ermittlung des Zusammenhangs bei Klausurnoten verschiedener Fächer. Ich möchte hier aber ein anderes konstruiertes Beispiel verwenden.

Für zehn Angestellte wurde mit einer Testarbeit sowohl ihre organisatorische Geschicklichkeit X als auch ihre Arbeitssorgfalt Y ermittelt. Es ergaben sich folgende Bewertungszahlen und Rangziffern:

Angestellter i	Organisatorisches Geschick		Arbeitssorgfalt	
	Punkte	Rang	Punkte	Rang
1	59	7	22	3
2	45	3	38	9
3	63	9	40	10
4	70	10	35	8
5	37	1	33	7
6	53	5	17	1
7	46	4	25	5
8	54	6	23	4
9	43	2	18	2
10	60	8	29	6

Der Rangkorrelationskoeffizient lautet $r_s = + 0{,}2848$.

Es ist also nur eine schwach ausgeprägt gleichsinnige Tendenz vorhanden.

Berechnet man anhand der vergebenen Punkte den Korrelationskoeffizienten nach Pearson, so erhält man $r = + 0{,}2188$.