

STATISTIK I – Übung 09

Spearman und Kendall

1 Kurze Wiederholung

Was sind Rangkorrelationskoeffizienten?

Liegen keine metrisch skalierten Daten vor, ist die Berechnung des Bravais-Pearson-Korrelationskoeffizienten (welcher zudem einen linearen Zusammenhang voraussetzt) nicht möglich. In diesem Fall (sowie auch in Fällen, in denen in metrischen Daten kein linearer Zusammenhang zu vermuten ist), können alternativ der Rangkorrelationskoeffizient nach Spearman sowie der Konkordanzkoeffizient nach Kendall berechnet werden. Beide setzen lediglich ordinalskalierte Daten voraus und geben die Stärke des monotonen Zusammenhangs zwischen zwei Variablen wieder, wobei in gleichsinnige (große Werte der einen Variablen gehen tendenziell mit großen Werten der anderen Variablen einher) und gegensinnige (große Werte der einen Variablen gehen tendenziell mit kleinen Werten der anderen Variablen – und umgekehrt – einher) zu unterscheiden ist. Liegt ein linearer Zusammenhang vor, existiert stets auch ein monotoner Zusammenhang – umgekehrt kann aber durchaus ein monotoner Zusammenhang vorliegen, ohne dass auch ein linearer Zusammenhang existiert.

Der Grundgedanke hinter beiden Koeffizienten beruht auf dem Umstand, dass sich sowohl ordinale als auch metrische Daten in eine natürliche Reihenfolge bringen, d.h. ordnen lassen. Untersucht man den Grad der Korrelation zwischen zwei Reihen solcher Daten, kann man somit eine der Datenreihen ordnen und anschließend prüfen, inwiefern sich die zweite Datenreihe „mitgeordnet“ hat. Liegt ein perfekter gleichsinniger Zusammenhang vor, so ist zu erwarten, dass sich die zweite Datenreihe perfekt mitordnet, d.h. (je nach gewählter Sortierung) entweder vom kleinsten zum größten oder vom größten zum kleinsten Datensatz. Im Falle eines perfekt gegensinnigen Zusammenhangs wäre dagegen zu erwarten, dass sich die zweite Datenreihe exakt entgegengesetzt zur ersten Datenreihe mit-sortiert. Alle anderen Fälle weichen mehr oder weniger stark von diesen beiden Sonderfällen ab.

Mit den Koeffizienten nach Spearman und Kendall werden wir nachfolgend zwei Wege kennenlernen, mit denen wir den Grad der Abweichung in der Mitsortierung der zweiten Datenreihe von den beiden Sonderfällen der perfekten gleichsinnigen und der perfekten gegensinnigen Korrelation auf verschiedene Arten ermitteln, und in einer leicht interpretierbaren Kennzahl ausdrücken können.

Rangkorrelationskoeffizient nach Spearman

Zur Bestimmung des Rangkorrelationskoeffizienten nach Spearman werden beide Datenreihen der Größe nach mit Rängen beziffert. Über die Summe der – im Falle eines perfekten gleichsinnigen Zusammenhangs nicht vorhandenen – quadrierten Differenzen zwischen den Rängen der ersten und der zweiten Datenreihe wird anschließend der Korrelationskoeffizient wie folgt berechnet:

$$rho = 1 - \frac{6 * \sum d_i^2}{(n^2 - 1) * n}$$

Die Vorgehensweise lässt sich am schnellsten anhand eines einfachen Beispiels erschließen:

Werte von x	Ränge von x	Werte von y	Ränge von y	Differenzen d	d ²
30	3	90	4	-1	1
20	2	70	2	0	0
40	4	100	5	-1	1
10	1	60	1	0	0
50	5	80	3	2	4

Als Summe der quadrierten Rangdifferenzen ergibt sich hier also 6. Eingesetzt (gemeinsam mit $n = 5$) in die Formel für den Rangkorrelationskoeffizienten nach Spearman ergibt sich folgender Wert:

$$rho = 1 - \frac{6 * \sum d_i^2}{(n^2 - 1) * n} = 1 - \frac{6 * 6}{(5^2 - 1) * 5} = 1 - \frac{36}{120} = 1 - 0,3 = 0,7$$

Der Wert für „Spearman's Rho“ liegt stets zwischen -1 und +1 und ist wie folgt zu interpretieren:

- > 0 -> gleichsinniger monotoner Zusammenhang
- ≈ 0 -> kein monotoner Zusammenhang
- < 0 -> gegensinniger monotoner Zusammenhang

Ein Problem für die Aussagekraft des Koeffizienten stellt das Auftreten sogenannter verbundener Ränge (auch Rangplatzbindungen oder ties) dar: Treten Werte mehrfach auf, bekommen diese einen Durchschnittsrang (berechnet als arithmetisches Mittel der zu vergebenden Ränge) zugewiesen.

Werte von x	Ränge von x
10	2
20	4,5
10	2
10	2
20	4,5

Je mehr verbundene Ränge existieren, umso geringer ist die Aussagekraft des Rangkorrelationskoeffizienten. Eine allgemeingültige Regel, ab welchem Anteil von verbundenen Rängen der Koeffizient nicht mehr berechnet werden sollte, existiert allerdings nicht, d.h. es bleibt eine subjektive Entscheidung. Eine Konstellation wie im obigen Beispiel dürfte allerdings in jedem Fall inakzeptabel sein.

Konkordanzkoeffizient nach Kendall

Auf dem gleichen gedanklichen Prinzip wie der Rangkorrelationskoeffizient nach Spearman basiert auch der Konkordanzkoeffizient nach Kendall. Für dessen Berechnung müssen ebenfalls zunächst die Ränge beider Variablenreihen gebildet werden. Anschließend werden die Daten nach den Rängen einer der beiden Datenreihen geordnet und wiederum überprüft, inwieweit sich die zweite Datenreihe „mitsortiert“ hat. Diese Überprüfung erfolgt anhand des paarweisen Vergleichs der Ränge der zweiten Datenreihe, wobei in konkordante (natürliche Reihenfolge der Ränge wird eingehalten) und diskordante Paare (natürliche Reihenfolge der Ränge wird nicht eingehalten) unterschieden wird.

Auch dieses Prinzip lässt sich am besten anhand des bereits bekannten Beispiels illustrieren:

Werte von x	Ränge von x	Werte von y	Ränge von y	Konkordante Paare	Diskordante Paare
10	1	60	1	4	0
20	2	70	2	3	0
30	3	90	4	1	1
40	4	100	5	0	1
50	5	80	3	//	//

In der ersten Zeile wird nun der y-Rang (1) mit den folgenden Rängen (2,4,5,3) verglichen. Dabei zeigt sich, dass auf die (1) nur größere Ränge folgen, die natürliche Ordnung also bei allen vier Paarvergleichen eingehalten wird. Für diese Zeile sind somit 4 konkordante und 0 diskordante Paare zu notieren.

Eine erste Abweichung zeigt sich in der dritten Zeile. Der Rang (4) ist hier mit den Rängen (5,3) zu vergleichen, wobei nur ein Paarvergleich konkordant ($4 < 5$), der andere jedoch diskordant ($4 > 3$) ausfällt. Insgesamt ergeben sich 8 konkordante (K) und 2 diskordante (D) Paare, die neben n (hier: 5) in die Formel für den Konkordanzkoeffizienten nach Kendall einzusetzen sind:

$$\tau = \frac{2 * (K - D)}{n * (n - 1)} = \frac{2 * (8 - 2)}{5 * (5 - 1)} = \frac{12}{20} = 0,6$$

Wie beim Rangkorrelationskoeffizienten nach Spearman ergibt sich auch hier ein Wert zwischen -1 und +1, der analog zu interpretieren ist. Obwohl beide Koeffizienten auf dem gleichen Prinzip des „Ordnen und Mitordnen“ basieren, fließen in den Konkordanzkoeffizienten nach Kendall weitaus mehr Informationen ein, da nicht nur einzelne Rangpaare (über die Differenzen), sondern alle Ränge miteinander abgeglichen werden (über die Ermittlung konkordanter und diskordanter Paare).

Ein Problem ergibt sich auch bei der Berechnung von Kendall durch verbundene Ränge, die in Paarvergleichen nicht mitgezählt werden. Im nachfolgenden, bereits bekannten Beispiel, ergeben sich etwa in 10 Paarvergleichen nur 6 Werte, da 4 Paarvergleiche mit identischen Werten ausscheiden.

Werte von x	Ränge von x	Konkordante Paare	Diskordante Paare
10	2	2	0
20	4,5	0	2
10	2	1	0
10	2	1	0
20	4,5	//	//

Korrelation und Kausalität

Zur korrekten inhaltlichen Interpretation von Korrelationen wurde in den vorangegangenen Übungen zum Bravais-Pearson-Korrelationskoeffizienten bereits einiges geschrieben, das auch für die Interpretation des Rangkorrelationskoeffizienten nach Spearman sowie des Konkordanzkoeffizienten nach Kendall Gültigkeit besitzt und bei der Bearbeitung entsprechender Aufgaben beachtet werden sollte.

2 Beispielrechnungen

Ein Kinobetreiber verändert (*ceteris paribus*) an 10 Tagen die Eintrittspreise für den Abendfilm und zeichnet die Besucherzahlen auf. Es ergibt sich die folgende Tabelle:

Abend Nr.	Preis pro Karte (in Euro)	Besucher/innen pro Abend
1	8,50	836
2	9,00	741
3	9,50	798
4	10,00	740
5	10,50	685
6	11,00	673
7	11,50	710
8	12,00	625
9	12,50	531
10	13,00	330

Bestimmung des Rangkorrelationskoeffizienten nach Spearman

Nr.	x	rg(x)	y	rg(y)	d	d ²
1	8,50	1	836	10	-9	81
2	9,00	2	741	8	-6	36
3	9,50	3	798	9	-6	36
4	10,00	4	740	7	-3	9
5	10,50	5	685	5	0	0
6	11,00	6	673	4	2	4
7	11,50	7	710	6	1	1
8	12,00	8	625	3	5	25
9	12,50	9	531	2	7	49
10	13,00	10	330	1	9	81

$$\sum d^2 = 322$$

$$rho = 1 - \frac{6 \cdot \sum d_i^2}{(n^2 - 1) \cdot n} = 1 - \frac{6 \cdot 322}{(10^2 - 1) \cdot 10} = 1 - \frac{1932}{990} = -0,9515$$

Bestimmung des Konkordanzkoeffizienten nach Kendall

Nr.	x	rg(x)	y	rg(y)	K	D
1	8,50	1	836	10	0	9
2	9,00	2	741	8	1	7
3	9,50	3	798	9	0	7
4	10,00	4	740	7	0	6
5	10,50	5	685	5	1	4
6	11,00	6	673	4	1	3
7	11,50	7	710	6	0	3
8	12,00	8	625	3	0	2
9	12,50	9	531	2	0	1
10	13,00	10	330	1	//	//

$$tau = \frac{2 \cdot (K - D)}{n \cdot (n - 1)} = \frac{2 \cdot (3 - 42)}{10 \cdot (10 - 1)} = \frac{-78}{90} = -0,8667$$

3 Übungsaufgaben (Lösungen folgen in der kommenden Woche)

Die Wiederholung des Kinopreis-Experiments in einem anderen Kino über die Dauer von 5 Tagen erbrachte folgende Ergebnisse:

Abend Nr.	Preis pro Karte (in Euro)	Besucher/innen pro Abend
1	8,50	527
2	9,00	489
3	9,50	512
4	10,00	315
5	10,50	308

a) Berechnen Sie den Rangkorrelationskoeffizienten nach Spearman.

b) Berechnen Sie den Konkordanzkoeffizienten nach Kendall.