 Zeitschrift für Soziologie, Jg. 7, Heft 3, August 1978, S. 267–272

Rangkorrelation bei soziologischen Normaldaten – Ein Ansatz zur Überwindung der Schwächen von Tau und Gamma

Gerhard Schulze
Universität Erlangen–Nürnberg
Institut für Soziologie und Sozialanthropologie

Rank correlation of typical sociological data – Proposals for an improvement of tau and gamma

A b s t r a c t : Usually, rank correlation is measured by tau-b, tau-c, or gamma. However, these coefficients are not fully adequate to most types of sociological data. Tau-b and tau-c cannot reach the extreme values of +1 and -1, given the marginal distribution of the variables, whereas gamma tends to overestimate weak associations. The coefficient „m", as it is presented in this paper, measures the correlation between ordinal variables without such deficiencies.

I n h a l t : Wenn Rangvariable aus sozialwissenschaftlichen Datensätzen miteinander korreliert werden, werden üblicherweise tau-b, tau-c oder gamma verwendet (wenn man parametrische Annahmen vermeiden möchte). Die tau-Koeffizienten haben jedoch den Mangel, daß sie im Normalfall je nach Variablenkombination veränderliche theoretische Extremgrenzen haben (statt konstant -1 und +1). Gamma hat zwar die konstanten Extremgrenzen -1 und +1, aber es tendiert zu einer Überschätzung des Zusammenhangs. Diese Probleme treten nicht auf bei dem im folgenden dargestellten Rangkorrelationskoeffizienten „m".

Selten wird in soziologischen Untersuchungen das ordinale Skalenniveau überschritten. Die gebräuchlichsten „reinen" Rangkorrelationskoeffizienten tau und gamma sind jedoch keineswegs für alle Rangvariablen in gleicher Weise geeignet; gerade bei solchen Daten, die für sozialwissenschaftliche Untersuchungen typisch sind, müssen Einwände gegen den Gebrauch von tau und gamma vorgebracht werden. SPEARMANs rho wird in die folgende Diskussion nicht einbezogen, weil es keine reine Rangkorrelationen zum Ausdruck bringt, sondern im Grunde parametrische Annahmen einschließt). Im folgenden werden Einwände gegen tau und gamma explizit gemacht und zum Ausgangspunkt einer Weiterentwicklung der Rangkorrelation für den speziellen Gebrauch der empirischen Sozialforschung genommen.

Die meisten Rangvariabelen in der empirischen Sozialforschung weisen nur „wenige" (etwa 3-6 Kategorien und sehr viele Rangbindungen innerhalb der einzelnen Kategorien auf, d.h.die einzelnen Ausprägungen der Variablen sind meist einer Vielzahl von Versuchspersonen gemeinsam (vgl. insbesondere Fragen nach Häufigkeiten und Intensitäten). Über die Verteilung der Fälle über die einzelnen Kategorien soziologischer Rang - variablen läßt sich überhaupt nichts aussagen – es gibt Rangvariable mit annähernder Gleichverteilung der Fälle ebenso wie solche mit einer linken, mittleren oder rechten Konzentration der Fälle. Unter diesen Umständen sind tau - Koeffizienten in aller Regel nicht mehr miteinander vergleichbar, um welche Version von tau es sich auch handeln mag.

Ursprünglich wurde KENDALLs tau für Rangvariable ohne Rangbindungen entwickelt, wo jede Versuchsperson ihren „eigenen" Rangplatz hat. Zwar gibt es auch „bindungskorrigierte" Versionen von tau (KENDALL 1970) für den Fall, daß die Kategorien der Variablen mehrfach besetzt sind (tau_b, tau_c), doch läßt sich zeigen, daß diese Koeffizienten nur in selten realisierten Spezialfällen die theoretischen Grenzen von + 1 und - 1 aufweisen: nur dann nämlich, wenn beide Variablen gleich viele Kategorien aufweisen und wenn sich bei gegebenen Randhäufigkeiten alle Fälle in einer der Hauptdiagonalen der bivariaten Datenmatrix anordnen lassen.Dies ist im Beispiel A der Fall; im Beispiel B ist nur eine Annäherung an die ideale montone Ordnung der Fälle innerhalb der bivariaten Datenmatrix möglich. Bei diesen Beispielen sind die Ranghäufigkeiten der Variablen gegeben; davon ausgehend wird versucht, die Häufigkeiten innerhalb der Matrix „so monoton wie möglich" zu verteilen. Im Beispiel B läßt sich keine idelae monotone Ordnung herstellen; unter diesen Um - ständen kann tau selbst dann nicht 1 werden, wenn sich die Variablen tatsächlich maximal monoton zueinander verhalten.

Beispiel A

10				10
	20			20
		20		20
			10	10
10	20	20	10	60

Beispiel B

5	15				20
		10	10	10	30
				10	10
5	15	10	10	20	60

Durch die Kategorienzahl und die Randverteilungen der beiden Variablen, die miteinander korreliert werden, ist von Anfang an festgelegt, ob tau (in welcher Version auch immer) die Grenzen von +1 und −1 überhaupt erreichen kann oder – was wesentlich häufiger der Fall ist – ob die theoretischen Extremgrenzen des Koeffizienten für ein spezifisches Variablenpaar „irgendwo" näher bei Null liegen (vgl. auch BENNINGHAUS 1976: 155 ff.). Es sind also nur solche tau-Koeffizienten vergleichbar, wo eine quadratische Datenmatrix zugrundeliegt und sich alle Fälle in den Hauptdiagonalen unterbringen lassen; alle anderen tau-Koeffizienten haben spezifische Extremgrenzen. Findet man bei zwei verschiedenen Variablenpaaren ein tau von jeweils gleicher Größe, so kann man deshalb in der Regel nicht schlußfolgern, daß der Zusammenhang in den beiden Fällen gleich stark sei.

Das von GOODMAN und KRUSKAL (1959) entwickelte Maß gamma, das sich bei der Korrelation von Rangvariablen steigender Beliebtheit erfreut, weist diesen Mangel nicht auf - es hat bei jeder beliebigen Kombination von Variablen die theoretischen Extremgrenzen +1 und −1. Gamma hat jedoch einen anderen Nachteil: Es kann zwischen den verschiedenen Graden linear - monotonen Zusammenhangs von Variablen nicht optimal differenzieren. Dies läßt sich am besten mit Hilfe folgender Vorstellung verdeutlichen: „Monotoner Zusammenhang" zwischen zwei Variablen ist ein Kontinuum, welches von einem Nullpunkt bis zu einem genau festliegenden Maximum reicht. Jedem Punkt auf diesem Kontinuum entspricht ein bestimmter Wert von gamma. Zwischen den Punkten auf dem Zusammenhangskontinuum und den Werten von gamma besteht jedoch keine lineare Entsprechung. Vielmehr steigt gamma in dem Bereich „schwacher" Zusammenhänge überproportional schnell an und weist immer geringere Steigerungsraten auf, je stärker der Zusammenhang wird. Gamma reagiert im Bereich schwacher Zusammenhänge zu sensibel, im Bereich starker Zusammenhänge zu wenig sensibel. Dies sei hier nur kurz angedeutet, es soll weiter unten noch ausführlicher behandelt werden.

Der im folgenden vorgestellte Rangkorrelationskoeffizient „m" (m als Symbol für „monotoner Zusammenhang") ist von den genannten Nachteilen frei. Er kann – im Gegensatz zu tau – in jedem Fall die Grenzen von +1 und −1 erreichen, und er steht – im Gegensatz zu gamma – in einem linearen Verhältnis zum Ausmaß des Zusammenhangs zwischen zwei Variablen.

Der Koeffizient „m" soll das Ausmaß des (positiven oder negativen) monotonen Zusammenhangs zwischen zwei Variablen zum Ausdruck bringen. Ein perfekter monotoner Zusammenhang ist dann gegeben, wenn die bivariate Datenmatrix quadratisch ist (die beiden Variablen also gleich viele Ausprägungsstufen aufweisen) und sämtliche Wertkombinationen in einer und nur einer Hauptdiagonalen konzentriert sind (wobei die eine Hauptdiagonale postitive Monotonie anzeigt, die andere Diagonale negative Monotonie). Bei weitaus den meisten Variablenkombinationen in der empirischen Sozialforschung ist jedoch dieses Ideal der perfekten Monotonie von Anfang an nicht erreichbar, und zwar aus zwei Gründen : (1) Die beiden Variablen weisen unterschiedliche Kategorienzahlen auf, so daß die bivariate Datenmatrix nicht quadratisch ist; (2) selbst bei quadratischen Datenmatrixen sind die Randverteilungen so beschaffen, daß kein Fall denkbar ist, wo alle Wertkombinationen in einer der Hauptdiagonalen konzentriert sind.

Für jede Variablenkombination existiert ein Maximum positiver Monotonie und ein Maximum negativer Monotonie. Diese Maxima bleiben in den meisten Fällen hinter dem Modell des perfekten monotonen Zusammenhanges mehr oder minder weit zurück. Die Grundidee des Koeffizienten m besteht nun darin, das tatsächlich in der Datenmatrix vorfindbare Ausmaß an monotonem Zusammenhang auf das theoretische Maximum an Monotonie zu beziehen, welches für jede Variableenkombination spezifisch ist:

$$m = \frac{\text{tatsächliches Ausmaß an Monotonie}}{\text{maximales Ausmaß an Monotonie}}$$

Zur Berechnung des tasächlichen Ausmaßes an Monotonie wird ebenso vorgegangen wie bei tau und gamma: Es wird die Differenz zwischen der Summe positiver monotoner Wertkombinationen (S_p) und der Summe negativer monotoner Wertkombinationen (S_n) gebildet (Logik und Berechnung von S_p und S_n sind etwa dargestellt bei FRÖHLICH/BECKER 1971: 505 ff., oder bei BENNINGHAUS 1976: 140 ff. Vgl. auch das BerechnungsbeispielS. 270 f.). Im Zähler des Quotienten steht also bei allen Koeffizienten tau, gamma und m dieselbe Größe. Jeweils unterschiedlich ist der Nenner definiert. Nur bei m, nicht bei den anderen Koeffizienten, ist die Bezugsgröße im Nenner bei jedem beliebigen Variablenpaar das maximale Ausmaß an (positiver oder negativer) Monotonie, welches für dieses Variablenpaar denkbar ist.

Das maximale Ausmaß für eine spezifische Variablenkombination kann errechnet werden, indem - ausgehend von den gegebenen Randhäufigkeiten- die Fälle innerhalb der Datenmatrix „so monoton wie möglich" angeordnet werden. Dabei gibt es zwei verschiedenen Lösungen: eine für positive Monotonie, und eine für negative Monotonie. Im folgenden Beispiel wird dies für die Kombination der Variablen X und Y durchgeführt:

	X1	X2	X3	X4	
Y1					69
Y2					70
Y3					86
Y4					152
Y5					136
	63	137	187	126	

Ausgangspunkt:Gegebene Randhäufigkeiten von X und Y

	X1	X2	X3	X4	
Y1	63	6			69
Y2		70			70
Y3		61	25		86
Y4			152		152
Y5			10	126	136
	63	137	187	126	

maximal positiv- monotone Anordnung der Fälle

	X1	X2	X3	X4	
Y1				69	69
Y2			13	57	70
Y3			86		86
Y4		64	88		152
Y5	63	73			136
	63	137	187	126	

maximal negativ - monotone Anordnung der Fälle

Das maximale Ausmaß an (positiver oder negativer) Monotonie für zwei gegebene Variblen kann errechnet werden als Differenz S_p - S_n, welche sich bei „möglichst monotoner" Anordnung der Fälle innerhalb der Datenmatrix ergibt. Bei einer maximal-positiv - monotonen Anordnung erreicht S_p ein Maximum und S_n wird Null; bei einer maximal negativ - monotonen Anordnung ist es umgekehrt.

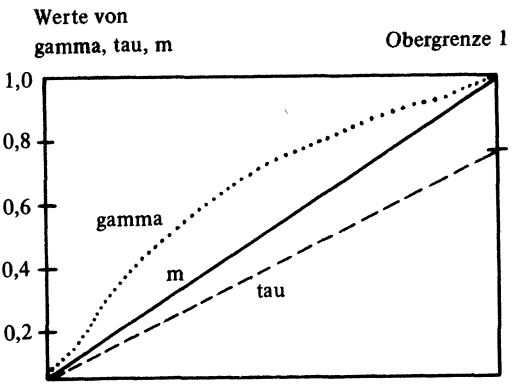
Der Koeffizient m ist also das tatsächliche Ausmaß an Monotonie (ausgedrückt als Differenz S_p - S_n für die empirisch vorgefundene Datenmatrix), bezogen auf das größtmögliche Asumaß an Monotonie für die gegebene Variablenkombination (ausgedrückt als absolute Differenz S_p- S_n für die monoton geordnete Datenmatrix). Ob bei der Errechnung des Nenners eine positiv - monotone oder eine negativ - monotone Anordnung der Datenmatrix anzulegen ist, kann entsprechend dem Vorzeichen des Zählers entschieden werden. Dieses Vorzeichen gibt darüber Auskunft, ob die Daten zweckmäßigerweise anhand eines Modells positiver oder negativer Monotonie zu beurteilen sind.

Bei den verschiedenen Versionen von tau ist die Bezugsgröße im Nenner fast immer „zu groß" – sie überstiegt in der Regel die maximalen Summendifferenzen, die sich bei möglichst monotonen Ordnungen der Datenmatrix ergeben. Damit ist es ausgeschlossen, daß die Extremgrenzen +1 und −1 erreicht werden können – tau-Koeffizienten haben unexplizierte Extremgrenzen, die „irgendwo" näher am Nullpunkt liegen; sie sind mithin nicht vergleichbar. Dies gilt insbesondere bei sozialwissenschaftlichen Normaldaten (nicht quadratische Datenmatritzen, unregelmässige Randhäufigkeiten).

Im Nenner des Koeffizienten gamma steht eine andere Größe: die Summe $S_p + S_n$. Dies hat die Konsequenz, daß es auch bei gegebenen Randverteilungen kein konstantes Bezugskriterium zur Beurteilung der tatsächlich vorgefundenen Monotonie gibt: Das Bezugskriterium $S_p + S_n$ ist variabel; es ist umso kleiner, je geringer der tatsächliche Zusammenhang zwischen den Variablen ist. Erst wenn der tatsächliche Zusammenhang maximal monoton ist, erreicht die Summe $S_p + S_n$ den Maximalwert, der bei der Berechnung von m konstant zugrundegelegt wird. Gamma hat also einen einprogrammierten variablen Auwertungseffekt, der im Bereich gemäßigter bis mittlerer Beziehungen besonders stark ist.

Man kann sich die Schwäche von gamma auch folgendermaßen verdeutlichen: der monotone Zusammenhang zwischen zwei Variablen ist ein Kontinuum, welches von Null bis zu einem Maximum reicht (positive oder negative Richtung des Zusammenhangs spielt für die folgenden Überlegungen keine Rolle). Dieses Kontinuum läßt sich bei Ordinalvariablen am besten durch die absolute Differenz $S_p - S_n$ abbilden, die ja auch im Zähler aller hier diskutierten Koeffizienten tau, gamma und m steht. Als Maß für die Stärke des Zusammenhangs ist dieses Kontinuum jedoch ungeeignet, weil seine Ausdehnung von einem Variablenpaar zum anderen variiert. Durch den (jeweils unterschiedlich definierten) Nenner der Koeffizienten tau, gamma und m werden die Werte auf dem Kontinuum des Zusammenhangs (Differenz $S_p - S_n$) transformiert. Dabei sollten zwei Bedingungen erfüllt werden:
(1) Das Zusammenhangskontinuum wird so transformiert, daß sich einheitlich die Schwankungsgrenzen von −1 und +1 ergeben;
(2) Koeffizient und Ausmaß des Zusammenhanges (abgebildet auf dem Kontinuum) stehen in linearer Beziehung: Variationen auf dem Kontinuum $S_p - S_n$ und Variationen des Koeffizienten stehen in gleicher Proportion.

Die erste Bedingung ist bei tau nicht erfüllt, die zweite fehlt bei gamma. Der Koeffizient m erfüllt beide Bedingungen. Im Diagramm weiter unten gibt die waagrechte Achse das Ausmaß des Zusammenhanges an, ausgedrückt durch die absolute Differenz $S_p - S_n$. Diese Differenz hat für jedes beliebige Variablenpaar ein feststehendes positives und negatives Maximum (im Diagramm wird der Einfachheit halber nur ein Maximum betrachtet). Die senkrechte Achse bezieht sich auf die Höhe der Koeffizienten tau, gamma und m. Die drei eingezeichneten Linien verdeutlichen das Verhältnis dieser Koeffizienten zum Kontinuum des Zusammenhangs: tau wächst zwar linear mit dem Ausmaß des Zusammenhangs an, erreicht jedoch typischerweise nicht die Grenze 1. Gamma erreicht zwar die Grenze 1, wächst jedoch nicht linear mit dem Ausmaß des Zusammenhangs an, sondern zunächst überproportional, dann unterproportional. Die Größe m erreicht die Grenze 1 und verhält sich zum Kontinuum des Zusammenhangs linear. Aus der Zeichnung ist auch die typische Rangfolge der Koeffizienten erkennbar: Gamma > m > tau (nur in den seltenen Fällen, wo eine quadratische Datenmatrix vorliegt und die Randverteilungen es nicht ausschließen, daß sämtliche Wertkombinationen in einer der Hauptdiagonalen liegen, ist der Wert von tau identisch mit dem Wert von m):



Kontinuum des monotonen Zusammenhangs zwischen Rangvariablen, ausgedrückt durch die absolute Differenz $S_p - S_n$

BERECHNUNG UND ZAHLENBEISPIEL

Die Berechnung des Koeffizienten m erfolgt nach der Formel

$$m = \frac{S_p - S_n}{S_{max}}$$

Ausgangspunkt der Berechnung ist die bivariate Datenmatrix. Die Berechnung läßt sich in drei

Schritte unterteilen, wovon die ersten beiden auch zur Routine bei der Berechnung von tau und gamma gehören:

1. Berechnung von S_p: Summe aller Produkte

absolute Häufigkeit in einer bestimmten Zelle der Datenmatrix	mal	Summe aller absoluten Häufigkeiten in den Zellen rechts und gleichzeitig unterhalb von den Ausgangszellen

2. Berechnung von S_n: Summe aller Produkte

absolute Häufigkeit in einer bestimmten Zelle der Datenmatrix	mal	Summe aller absoluten Häufigkeiten in den Zellen links und gleichzeitig unterhalb von der Ausgangszelle

3. Berechnung von S_{max}:

Zur Berechnung von S_{max} muß zunächst die Datenmatrix ausgehend von den gegebenen Randhäufigkeiten maximal monoton geordnet werden (in der weiter oben angedeuteten Weise). Ob eine positiv-monotone oder negativ-monotone Ordnung herzustellen ist, entscheidet sich nach dem Vorzeichen von $S_p - S_n$.

Nun kann S_{max} nach der folgenden Formel errechnet werden:

$$S_{max} = | S'_p - S'_n |$$

Die Größen S'_p und S'_n beziehen sich auf die neugebildete Matrix. Sie werden in gleicher Weise errechnet wie dei Größen S_p und S_n. In der monoton geordneten Matrix muß immer eine der beiden Größen den Wert Null haben.

ZAHLENBEISPIEL:

Ausgangsmatrix	X1	X2	X3	X4	
Y1	50	20	10	10	90
Y2	20	30	60	50	160
Y3	10	20	50	50	130
	80	70	120	110	380

1. $S_p = 50\,(30 + 60 + 50 + 20 + 50 + 50) + 20\,(60 + 50 + 50 + 50) + 10\,(50 + 50) + 20\,(20 + 50 + 50) + 30\,(50 + 50) + 60\,(50) = 26600$

2. $S_n = 20\,(20 + 10) + 10\,(20 + 30 + 10 + 20 + 10\,(20 + 30 + 60 + 10 + 20 + 50) + 30\,(10) + 60\,(10 + 20) + 50\,(10 + 20 + 50) = 9400$

3. Die Differenz $S_p - S_n$ ist positiv, also wird die Matrix zur maximalen positiven Ordnung umorganisiert:

	X1	X2	X3	X4	
Y1	80	10			90
Y2		60	100		160
Y3			20	110	130
	80	70	120	110	

$$S_p' = 80\,(60 + 100 + 20 + 110) + (100 + 20 + 110) + 60\,(20 + 110) + 100\,(110) = 44\,300$$

$$S_n' = 0$$

$$S_{max} = | S_p' - S_n' | = 44\,300$$

Daraus Berechnung von m:

$$m = \frac{S_p - S_n}{S_{max}} = \frac{26600 - 9400}{44300} = 0{,}39$$

Für dieselbe Matrix errechnen sich folgende Vergleichswerte von tau und gamma: tau-b = 0,34 tau-c = 0,35 gamma = 0,48

Bei den gegebenen Randhäufigkeiten haben die Koeffizienten folgende Maximalwerte im positiven Bereich:

tau-b = 0,89
tau-c = 0,92
gamma = 1,0
m = 1,0

Wegen der systematischen Unterschätzungstendenzen von tau auf der einen Seite und der systematischen Überschätzungstendenz von gamma auf der anderen Seite klaffen tau und gamma erheblich auseinander. Der Koeffizient m liegt zwischen diesen Extremen. Er ist sinnvoll interpretierbar als ein Punkt auf dem Kontinuum monotonen Zusammenhanges zwischen Rangvariablen mit den Grenzen +1 und −1.

Literatur

BENNINGHAUS, H., 1976: Deskriptive Statistik. Stuttgart.

FRÖHLICH, W. D., J. BECKER, 1971: Forschungsstatistik. Bonn.

GODMAN, L. A., W. H. KRUSKAL, 1959: Measures of Association for Cross-Classifications. Journal of American Statistical Association 54, 123 ff.

KENDALL, M. G., 1970: Rank Correlation Methods. London (4. Aufl.).

Anschrift des Verfassers:
Dr. habil. GERHARD SCHULZE
Hauptstraße 4
8521 Bubenreuth