Der Kappa-Koeffizient

– Artikel Nr. 23 der Statistik-Serie in der DMW –

The kappa coefficient

Autoren U. Grouven [1] R. Bender [1] A. Ziegler [2] S. Lange [1]

Institut
[1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln
[2] Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck

Übereinstimmung von Bewertungen

Der Zuverlässigkeit von Beobachtungen und Bewertungen kommt in der medizinischen Forschung eine große Bedeutung zu. Häufig ist es von Interesse, die Übereinstimmung mehrerer Bewertungen ein und desselben Sachverhaltes zu untersuchen, um Fehlerquellen durch Variabilität der Messungen zu identifizieren und zu quantifizieren. Dabei sind prinzipiell zwei Fälle zu unterscheiden: **1**) die Beurteilung durch unterschiedliche Bewerter und **2**) die wiederholte Beurteilung durch einen einzelnen Bewerter. Im ersten Fall spricht man auch von ***inter***rater agreement, im zweiten Fall von ***intra***rater agreement. Beispiele hierfür sind die Befundung eines Röntgenbildes durch zwei verschiedene Radiologen oder durch einen Radiologen zu unterschiedlichen Zeitpunkten, oder die wiederholte Durchführung eines diagnostischen Tests in unterschiedlichen Labors oder eine wiederholte Testdurchführung in einem einzelnen Labor. Die Untersuchung der Übereinstimmung – oder Konkordanz – gibt Aufschluss über Stabilität und Zuverlässigkeit der Diagnose bzw. des Testverfahrens.

Das meistverwendete Maß zur Bewertung der Übereinstimmungsgüte bei Vorliegen von kategoriellen Merkmalen ist Cohens Kappa-Koeffizient [6], der im Folgenden näher beschrieben wird. Die Methoden zur Untersuchung der Übereinstimmungsgüte von stetigen Daten werden in einem anderen Artikel behandelt [10].

Kappa-Koeffizient

Zunächst soll der einfachste Fall von zwei Bewertern, die N Patienten basierend auf einem geeigneten Kriterium in eine von zwei möglichen Kategorien „krank“ oder „gesund“ klassifizieren, betrachtet werden. Die beobachteten Häufigkeiten lassen sich in Form einer Vierfeldertafel – auch Übereinstimmungsmatrix oder Klassifikationstabelle genannt – darstellen (**Tab. 1**). Diese Tabelle sieht formal genau so aus wie die Vierfeldertafeln, die zum Vergleich von zwei Gruppen oder zur Untersuchung diagnostischer Tests verwendet werden [3]. Die Fragestellung ist hier aber eine andere, so dass hier auch andere Methoden benötigt werden.

Tab. **1** Übereinstimmungsmatrix bei zwei Bewertern und zwei möglichen Kategorien.

Bewerter 1	**Bewerter 2**		
	krank	gesund	gesamt
krank	a	b	a + b
gesund	c	d	c + d
gesamt	a + c	b + d	N = a + b + c + d

Übereinstimmende Bewertungen finden sich in der Hauptdiagonalen (a und d), abweichende Beurteilungen in den übrigen Zellen (b und c). Es ist nun nahe liegend, als Maßzahl für die Übereinstimmungsgüte den relativen Anteil der übereinstimmenden Messungen an der Gesamtzahl N, also $p_o = (a+d)/N$ zu betrachten. Dabei ist jedoch zu beachten, dass ein gewisses Maß an Übereinstimmung auch dann zu erwarten ist, wenn die beiden Bewerter rein zufällig urteilen würden. Die Idee bei der Berechnung des Kappa-Koeffizienten ist es nun, den Anteil rein zufälliger Übereinstimmung aus dem beobachteten Anteil von Übereinstimmungen „herauszurechnen“. Seien nun $p_{1,krank} = (a+b)/N$ und $p_{1,gesund} = (c+d)/N$ die Anteile von Bewerter 1 als „krank“ bzw. „gesund“ eingestufter Patienten und $p_{2,krank} = (a+c)/N$ und $p_{2,gesund} = (b+d)/N$ die entsprechenden Anteile von Bewerter 2 als „krank“ bzw. „gesund“ bewerteter Patienten. Unter der Annahme der Unabhängigkeit der Bewerter und bei gegebenen Randhäufigkeiten berechnet sich die geschätzte Wahrscheinlichkeit einer zufälligen Übereinstimmung

Schlüsselwörter
- Kappa
- Übereinstimmung
- Vierfeldertafel
- Konfidenzintervall

Key words
- Kappa
- Agreement
- 2 by 2 table
- Confidence interval

Bibliografie
DOI 10.1055/s-2007-959046
Dtsch Med Wochenschr 2007; 132: e65–e68 · © Georg Thieme Verlag Stuttgart · New York · ISSN 0012-0472

Korrespondenz
Privatdozent Dr. rer. biol. hum. Ralf Bender
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail Ralf.Bender@iqwig.de

stimmung in den Kategorien „krank“ und „gesund“ als Produkt der entsprechenden Anteile, also $p_{1,krank} \times p_{2,krank}$ und $p_{1,gesund} \times p_{2,gesund}$. Der Gesamtanteil zufälliger Übereinstimmungen ergibt sich dann durch Aufsummieren der Produkte als $p_e = (p_{1,krank} \times p_{2,krank}) + (p_{1,gesund} \times p_{2,gesund})$.

Der Kappa-Koeffizient ist definiert als

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Der Zähler ist die Differenz der Anteile von tatsächlich beobachteter und aufgrund von Zufall erwarteter Übereinstimmung, d. h. der Anteil von Übereinstimmungen, der über den Zufall hinaus geht. Der Nenner dient der Standardisierung. Der maximal erreichbare Wert von p_o ist 1 (bei völliger Übereinstimmung), in diesem Fall ist $\kappa = 1$. Ein Wert von 0 bedeutet einen nicht mehr als zufallbedingten Grad an Übereinstimmung. Negative Werte sind theoretisch auch möglich und implizieren eine schlechtere Übereinstimmung als zufällig zu erwarten ist, was jedoch in der Praxis selten auftritt und als lediglich zufallsbedingte Übereinstimmung interpretiert werden sollte [4]. Formeln zur Berechnung des zugehörigen Standardfehlers und daraus abgeleiteter Konfidenzintervalle findet man z. B. bei Fleiss et al. [9].

Wie sind nun konkrete Werte von κ, z. B. 0.63, zu interpretieren? Allgemein gültige Aussagen sind nicht möglich, sondern immer von der konkreten Situation abhängig. In der Literatur werden jedoch Richtwerte angegeben (z. B. [1], **Tab. 2**).

Das Prinzip bei der Berechnung von p_o, p_e und κ bleibt auch gültig bei Vorliegen von mehr als zwei Kategorien. Dies soll anhand des folgenden Rechenbeispiels veranschaulicht werden.

Beispiel

Tab. 3 zeigt die Bewertung von 118 Proben, die von 2 Pathologen unabhängig voneinander auf einer vierstufigen Skala mit Ausprägungen 1 = „negativ“, 2 = „atypische schuppenartige Hyperplasie“, 3 = „Karzinom in situ“ und 4 = „schuppenartiges oder invasives Karzinom“ vorgenommen wurde [2].

Der Anteil der Übereinstimmungen ergibt sich durch Aufsummieren der Hauptdiagonalelemente und anschließende Division durch die Gesamtzahl der Patienten, d. h.

$$p_o = \frac{22 + 7 + 36 + 10}{118} = 0{,}636$$

Der Anteil von Übereinstimmungen, der bei unabhängiger zufälliger Bewertung zu erwarten ist, beträgt

$$p_e = \left(\frac{26}{118} \times \frac{27}{118}\right) + \left(\frac{26}{118} \times \frac{12}{118}\right) + \left(\frac{38}{118} \times \frac{69}{118}\right) + \left(\frac{28}{118} \times \frac{10}{118}\right) = 0{,}281$$

Für den Kappa-Koeffizienten ergibt sich somit:

$$\kappa = \frac{0{,}636 - 0{,}281}{1 - 0{,}281} = 0{,}493$$

Tab. **2** Richtwerte zur Interpretation von κ.

Wert von κ	Stärke der Übereinstimmung
< 0,20	schwach
0,21 – 0,40	leicht
0,41 – 0,60	mittelmäßig
0,61 – 0,80	gut
0,81 – 1,00	sehr gut
nach Altman [1]	

Tab. **3** Übereinstimmungsmatrix von 2 Pathologen bei der Klassifikation von 118 Proben in 4 mögliche Kategorien.

Pathologe 1	Pathologe 2				
	1	2	3	4	gesamt
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
gesamt	27	12	69	10	118
nach Agresti [2]					

Der Standardfehler beträgt SE(κ) = 0,057 und das 95%-Konfidenzintervall umfasst den Bereich von 0,382 bis 0,604. Gemäß den Richtwerten in **Tab. 2** liegt somit eine mittelmäßige Übereinstimmung zwischen den Bewertungen der beiden Pathologen vor.

Erweiterungen des Kappa-Koeffizienten

Es existiert eine Reihe von Modifikationen und Erweiterungen des Kappa-Koeffizienten. Eine mögliche Erweiterung stellt das gewichtete Kappa κ_w dar [7]. Haben die Bewertungskategorien ordinales Messniveau, so ist es häufig der Fall, dass Abweichungen um mehrere Kategorien schwerer wiegen als Abweichungen um lediglich eine Kategorie. Um dies bei der Berechnung der Übereinstimmungsgüte zu berücksichtigen, wird die herkömmliche Formel zur Berechnung von κ durch Berücksichtigung geeigneter Gewichte zwischen 0 und 1 modifiziert. Im obigen Beispiel liegen ordinale Bewertungskategorien vor und für das gewichtete Kappa errechnet sich ein Wert von $\kappa_w = 0{,}649$. In der Literatur werden verschiedene Gewichte vorgeschlagen; hier wurden die Gewichte nach Cicchetti-Fleiss verwendet [9]. Zu beachten ist, dass der Wert von κ_w von der konkreten Wahl der Gewichte abhängt [4].

Der gewichtete Kappa-Koeffizient ist gewöhnlich größer als das ungewichtete Kappa, da Abweichungen in benachbarte Kategorien in der Regel häufiger vorkommen als Abweichungen um mehrere Kategorien [1].

Eine weitere Verallgemeinerung stellt die Situation dar, wenn mehr als 2 Bewertungen vorliegen und deren Übereinstimmung überprüft werden soll. Entsprechende Auswerteverfahren sind z. B. in [9] beschrieben.

Anmerkungen zur Anwendung und Interpretation von κ

Bei der Anwendung und Interpretation des Kappa-Koeffizienten sind eine Reihe von Aspekten zu beachten. So ist es möglich, dass völlig unterschiedliche Datenkonstellationen zu identischen Werten von Kappa führen. Zum besseren Verständnis ist es sinnvoll, sich einige Sachverhalte und Zusammenhänge vor Augen zu führen. Zum einen ist der Wert von Kappa abhängig

Tab. **4** Bewertung von 100 Röntgenbildern durch 2 Radiologen.

	Radiologe A	Radiologe B		
		positiv	negativ	gesamt
(a)	positiv	40	15	55
	negativ	15	30	45
	gesamt	55	45	100
(b)	positiv	65	15	80
	negativ	15	5	20
	gesamt	80	20	100
(c)	positiv	35	20	55
	negativ	10	35	45
	gesamt	45	55	100

von der Anzahl der Klassifikationskategorien. Bei Vorliegen von mehreren Kategorien wird eine übereinstimmende Klassifikation automatisch schwieriger und der Wert von Kappa wird tendenziell kleiner [1].

Bei näherer Betrachtung der Formel zur Berechnung des Kappa-Koeffizienten wird deutlich, dass der Wert von Kappa von der Größe des zufällig erwarteten Anteils p_e abhängt. Je größer p_e desto kleiner wird der Wert von Kappa. Das bedeutet, dass bei gegebener beobachteter Übereinstimmung p_o unterschiedliche Werte für Kappa resultieren können, je nach Größe von p_e. Der Wert von p_e ist abhängig von der „Prävalenz" des betrachteten Merkmals, d.h. von der Verteilung der Randhäufigkeiten. Ein maximaler Wert für Kappa von 1 (d. h. komplette Übereinstimmung der Bewertungen) ist nur möglich, wenn die Verteilungen der Randhäufigkeiten gleich sind (d. h. a+b = a+c und c+d = b+d in **Tab. 1**), ansonsten ergeben sich automatisch Werte außerhalb der Hauptdiagonalen.

Die Abhängigkeit des Kappa-Koeffizienten von der Verteilung der Randhäufigkeiten lässt sich an folgendem hypothetischen Beispiel weiter veranschaulichen. 100 Röntgenaufnahmen werden von zwei Radiologen hinsichtlich des Vorliegens eines Befundes als „positiv" oder „negativ" bewertet. **Tab. 4** zeigt drei mögliche Übereinstimmungsmatrizen.

Der Anteil der beobachteten Übereinstimmungen beträgt in allen Fällen $p_o = 0{,}70$ ([40 + 30]/100 in Fall (a), [65 + 5]/100 in Fall (b) und [35 + 35]/100 in Fall (c)). Die drei Beispiele unterscheiden sich jedoch in der Verteilung ihrer Randhäufigkeiten. Im Fall (a) liegt eine relativ balancierte, bei beiden Radiologen gleiche Verteilung der Randhäufigkeiten vor. Der Wert für Kappa beträgt hier $\kappa = 0{,}39$. Im Fall (b) sind die Verteilungen der Randhäufigkeiten ebenfalls symmetrisch zwischen den Radiologen, jedoch stark unbalanciert bzgl. der beiden Kategorien. Eine positive Bewertung wird von beiden Radiologen deutlich häufiger vorgenommen als eine negative Bewertung (80% positive, 20% negative Bewertungen). Trotz gleicher beobachteter Übereinstimmung von 70% im Vergleich zu Fall (a) sinkt der Wert von Kappa auf $\kappa = 0{,}06$. Dieses zunächst paradox anmutende Ergebnis lässt sich dadurch erklären, dass in Situation (b) bei beiden Radiologen der Anteil der positiven Bewertungen deutlich vorherrscht. Daher sind unterschiedliche Bewertungen kaum möglich und es bleibt nur wenig Spielraum für übereinstimmende Bewertungen, die über den erwarteten Anteil an zufälligen Übereinstimmungen hinausgehen [8]. In Fall (c) sind die generellen Bewertungen der Radiologen unsymmetrisch, d. h. es liegt ein Bias vor. Radiologe A bewertet 55% aller Fälle als positiv und 45% als negativ, bei Radiologe B ist es genau umgekehrt. Der Wert von Kappa ist in diesem Fall $\kappa = 0{,}41$ und damit größer als im Fall (a), wo die Bewertungshäufigkeiten für die unterschiedlichen Kategorien bei beiden Radiologen identisch sind. D. h. der Wert von Kappa ist im Fall (c) größer als im Fall (a), obwohl die generellen Bewertungshäufigkeiten der Bewerter im Fall (c) stärker voneinander abweichen als im Fall (a). Auch dieses scheinbar paradoxe Ergebnis erklärt sich durch die Zufallskorrektur bei der Berechnung des Kappa-Koeffizienten. Eine zufällige Übereinstimmung der Bewerter wird schwieriger bei asymmetrischer Verteilung der Randhäufigkeiten, dadurch wird der Korrekturfaktor p_e kleiner und der Wert von Kappa somit größer [8].

Tab. **5** Übersetzung wichtiger englischer Begriffe (deutsch – englisch).

Übereinstimmung zwischen Bewertern	interrater agreement
Übereinstimmung der Beurteilungen eines Bewerters	intrarater agreement
Konkordanz	concordance
Kappa-Koeffizient	kappa coefficient
Übereinstimmungsmatrix	agreement matrix
zufallskorrigiert	chance-corrected
Prävalenz	prevalence
gewichtetes Kappa	weighted kappa
Verzerrung	bias

Nähere Einzelheiten zu Modifikationen, Anwendungen und Eigenschaften des Kappa-Koeffizienten lassen sich in einer Reihe von Übersichtsarbeiten nachlesen [5, 11, 12, 13].

Bei der Interpretation von Studienergebnissen ist es von größter Wichtigkeit, nicht nur den Wert des Kappa-Koeffizienten, sondern auch die dazugehörige Klassifikationstabelle zu kennen. Dies ist insbesondere auch wichtig bei dem Vergleich von Kappa-Koeffizienten aus verschiedenen Studien. Zudem ermöglicht die Kenntnis der Klassifikationstabelle eine differenzierte Bewertung von Übereinstimmungen in den unterschiedlichen Kategorien.

Schließlich sei noch angemerkt, dass Übereinstimmung nicht gleichbedeutend ist mit Assoziation oder Korrelation. Eine hohe Übereinstimmung impliziert eine hohe Assoziation, aber der umgekehrte Fall trifft nicht notwendigerweise zu. Wenn z. B. in der Situation aus dem oben geschilderten Beispiel Pathologe 1 die Proben durchweg eine Kategorie höher einstuft als Pathologe 2, so sind die Bewertungen hoch korreliert, aber die Übereinstimmung der Bewertungen ist niedrig. Aus diesem Grund stellen Korrelationsmaße oder übliche χ^2-Tests keine geeigneten Auswertemethoden für die vorliegende Datensituation dar. Trotz der beschriebenen Einschränkungen stellt der Kappa-Koeffizient nach Cohen den „Gold Standard" zur Bewertung der Übereinstimmungsgüte bei kategoriellen Daten dar. Die englischen Übersetzungen der verwendeten Begriffe zeigt **Tab. 5**.

kurzgefasst

Die meistverwendete statistische Methode zur Auswertung der Übereinstimmung zwischen unterschiedlichen Bewertern oder zwischen wiederholten Beurteilungen eines Bewerters bezüglich eines kategoriellen Merkmals ist Cohens Kappa-Koeffizient. Cohens Kappa misst den zufallskorrigierten Anteil übereinstimmender Bewertungen.

Literatur

1 *Altman DG.* Practical Statistics for Medical Research. Chapman & Hall/CRC, Boca Raton, 1991
2 *Agresti A.* Categorical Data Analysis. Second edition. Wiley, 2002
3 *Bender R, Lange S.* Die Vierfeldertafel. Dtsch Med Wochenschr 2007; 132: e12–e14
4 *Brennan P, Silman A.* Statistical methods for assessing observer variability in clinical measures. BMJ 1992; 304: 1491–1494
5 *Byrt T, Bishop J, Carlin JB.* Bias, prevalence and kappa. Journal of Clincal Epidemiology 1993; 46: 423–424
6 *Cohen J.* A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960; 20: 37–46
7 *Cohen J.* Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70: 213–220
8 *Feinstein AR, Cicchetti DV.* High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol 1990; 43: 543–549
9 *Fleiss JL, Levin B, Paik MC.* Statistical Methods for Rates and Proportions. Third edition (Hrsg). Wiley, 2003
10 *Grouven U, Bender R, Ziegler A, Lange S.* Vergleich von Messmethoden. Dtsch Med Wochenschr 2007; 132: e69–e73
11 *Kraemer HC, Bloch AD.* Kappa coefficients in epidemiology. An appraisal of a reappraisal. J Clin Epidemiol 1988; 41: 959–968
12 *Maclure M, Willett WC.* Misinterpretation and misuse of the kappa coefficient. Am J Epidemiol 1987; 126: 161–169
13 *Thompson WD, Walter SD.* A reappraisal of the kappa coefficient. J Clin Epidemiol 1988; 41: 969–970