

Zusammenhangsanalysen nominaler Daten (1)

Liegen nur nominalskalierte Merkmale vor, können kategoriale Zusammenhänge zwischen den Variablen mittels **Kreuztabellen** der absoluten oder relativen Häufigkeiten analysiert werden.

Hyothetisches Beispiel: *Perfekte Unabhängigkeit* zwischen Geschlecht und Studienfach

	Sprach- & Kulturwiss.	Rechts-, Wirtsch.- & Sozialwiss.	Mathem. & Naturwiss.	total
weiblich	18	36	18	72
männlich	27	54	27	108
total	45	90	45	180

- 25% (18/72) der *Frauen* präferieren Sprach- u. Kulturwissenschaften, 50% (36/72) Rechts-, Wirtschafts- u. Sozialwissenschaften, 25% (18/72) Mathematik u. Naturwissenschaften,
 - 25% (27/108) der *Männer* präferieren Sprach- u. Kulturwissenschaften, 50% (54/108) Rechts-, Wirtschafts- u. Sozialwissenschaften, 25% (27/108) Mathematik u. Naturwissenschaften,
 - 40% aller *Sprach- u. Kulturwissenschaftsstudenten* sind weiblich (18/45), 60% sind männlich (27/45),
 - 40% aller *Rechts-, Wirtschafts- u. Sozialwissenschaftsstudenten* sind weiblich (36/90), 60% sind männlich (54/90),
 - 40% aller *Mathematik- u. Naturwissenschaftsstudenten* sind weiblich (72/180), 60% männlich (108/180).
- In der Tabelle gibt es *keinen* Zusammenhang zwischen Geschlecht und Studienfachpräferenz.**

Zusammenhangsanalysen nominaler Daten (2)

Fiktives Beispiel: Zusammenhang zwischen Geschlecht und Studienfachpräferenz

Fett: <i>Kursiv:</i> erwartet	Sprach- & Kulturwiss.	Rechts-, Wirtsch.- & Sozialwiss.	Mathem. & Naturwiss.	total
weiblich	28 (18)	31 (36)	13 (18)	72
männlich	17 (27)	59 (54)	32 (27)	108
total	45	90	45	180

Berechnungen:

Welche Häufigkeit würde man bei Unabhängigkeit erwarten?

$$f_e = \frac{\text{Zeilensumme} \cdot \text{Spaltensumme}}{N}$$

Diskrepanz-Statistik (Fit-Maß):

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
28	18	10	100	5.56
31	36	-5	25	0.69
13	18	-5	25	1.39
17	27	-10	100	3.70
59	54	5	25	0.46
32	27	5	25	0.93
				$\chi^2 = 12.73$

Zusammenhangsanalysen nominaler Daten (3)

Ähnlich wie die Kovarianz als Maß der Enge des Zusammenhangs wenig geeignet ist, eignet sich auch der **Chi²**-Wert nicht gut als Maßzahl des Zusammenhangs, da sein Wert nach oben unbeschränkt ist. Er sagt nur, **ob** ein Zusammenhang besteht, **nicht wie groß** er ist: **Chi²** hängt von der Stichprobengröße und Anzahl der Zellen der Kreuztabelle (Kontingenztafel) ab.

Ein klassisches Assoziationsmaß nominaler Daten ist der **Kontingenzkoeffizient C**:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \quad \text{sein Maximalwert ist} \quad C_{\max} = \sqrt{\frac{k-1}{k}} \quad \text{mit } k = \text{Minimum von Zeilen und Spaltenzahl}$$

Ein besseres Maß, dessen maximaler Wert nicht wie C von der Zeilen- oder Spaltenzahl abhängt und unabhängig von der Tabellengröße zwischen 0 und 1 liegt, ist

Cramérs V:

$$\text{Cramérs } V = \sqrt{\frac{\chi^2}{N(k-1)}} \quad \text{mit } k = \text{Minimum von Zeilen und Spaltenzahl}$$

Anhand des oben gezeigten **Beispiels** ist der Zusammenhang zwischen Geschlecht und Studienfachpräferenz:

$$C = \sqrt{\frac{12.73}{180 + 12.73}} = 0.257 \quad \text{und} \quad \text{Cramérs } V = \sqrt{\frac{12.73}{180 \cdot (2-1)}} = 0.266$$

Zusammenhangsanalysen nominaler Daten (4)

Sind die beide Variablen **natürlich dichotom** (z.B. Straffälligkeit und Geschlecht) kann ihre gemeinsame Verteilung in einer 2 x 2 - Kreuztabelle (Vierfeldertafel) dargestellt werden. Als Assoziationsmaß kann der **phi-Koeffizient** berechnet werden:

Vierfeldertafel:

	-X	X	Σ
-Y	a = n ₁₁	b = n ₁₂	(a + b)
Y	c = n ₂₁	d = n ₂₂	(c + d)
Σ	(a + c)	(b + d)	N

Formel:

$$\varphi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+c) \cdot (b+d) \cdot (a+b) \cdot (c+d)}} = r_{xy}$$

Der **phi-Koeffizient** ist **identisch mit der Produkt-Moment-Korrelation** beliebig kodierter dichotomer Daten. Je stärker die Häufigkeiten einer der Diagonalen (a, d) oder (c, b) dominieren, um so mehr weicht er von Null ab.

Der Betrag des **phi-Koeffizienten** ist gleich Cramérs V (das eine Verallgemeinerung des **phi-Koeffizienten** für beliebige Zeilen- und Spaltenanzahl darstellt) und lässt sich damit auch aus dem **Chi²**-Wert der Vierfeldertafel berechnen (das Vorzeichen muss dann aus der Besetzung der Diagonalen bestimmt werden). **Im Fall einer Vierfeldertafel gilt:**

$$|\varphi| = \text{Cramérs } V = \sqrt{\frac{\chi^2}{N}}$$

Zusammenhangsanalysen nominaler Daten (5)

Beispiel: Laut Strafverfolgungsstatistik für das Jahr 2003 wurden 1 461 weibliche und 91 251 männliche Jugendliche wegen Straftaten gegen das Leben und wegen Körperverletzungsdelikten verurteilt. Im gleichen Jahr lebten in der Bundesrepublik 1871.2 Tausend weibliche und 1975.0 Tausend männliche Jugendliche.

Wie hoch war 2003 der Zusammenhang von Geschlecht und Verurteilung wegen Straftaten gegen das Leben und Körperverletzungsdelikten bei Jugendlichen?

- nicht verurteilte weibl. Jugendl. (in Tausend) = $1871.2 - 1.461 = 1869.739$
- nicht verurteilte männl. Jugendl. (in Tausend) = $1975.0 - 91.251 = 1883.749$

Vierfeldertafel:

	weiblich	männlich	Σ
nicht verurteilt	1869.739	1883.749	3753.488
verurteilt	1.461	91.251	92.712
Σ	1871.200	1975.000	3846.200

Berechnung:

$$\varphi = \frac{1869.739 \cdot 91.251 - 1883.749 \cdot 1.461}{\sqrt{1871.2 \cdot 1975.0 \cdot 3753.488 \cdot 92.712}} = 0.148$$

- verurteilte weibliche Jugendliche: 0.08% ($100 \cdot 1.461 / 1871.2$)
- verurteilte männliche Jugendliche: 4.62% ($100 \cdot 91.251 / 1975.0$)

Weitere Korrelationskoeffizienten (1)

– Exkurs: Sensitivität, Spezifität, und Odds-Ratio (I) –

Bei zwei natürlich dichotomen Variablen ist der φ -Koeffizient (gleich der Produkt-Moment-Korrelation) ein geeignetes Assoziationsmaß. Sind die Merkmale jedoch **künstlich dichotom**, kann der φ -Koeffizient in die Irre führen.

Männer

	gesund	krank	Σ
negativ	518	117	635
positiv	182	183	365
Σ	700	300	1000

$$r = \varphi = .33$$

Qualität des Tests:

Sensitivität (Rate korrekt Positiver):
 $h(\text{positiv} | \text{krank}) = 183 / 300 = .61$

Spezifität (Rate korrekt Negativer):
 $P(\text{negativ} | \text{gesund}) = 518 / 700 = .74$

Frauen

	gesund	krank	Σ
negativ	666	39	705
positiv	234	61	295
Σ	900	100	1000

$$r = \varphi = .23$$

Sensitivität (Rate korrekt Positiver):
 $P(\text{positiv} | \text{krank}) = 61 / 100 = .61$

Spezifität (Rate korrekt Negativer):
 $P(\text{negativ} | \text{gesund}) = 666 / 900 = .74$

Die Qualität des Tests ist bei Männern und Frauen gleich! Der Unterschied der Korrelationen von Test und Krankheit (φ) ist hier durch die **unterschiedlichen Basisraten** (Auftrittswahrscheinlichkeit der Erkrankung) bei Männern ($300/1000 = 0.30$) und Frauen ($100/1000 = 0.10$) bedingt.

Weitere Korrelationskoeffizienten (2)

– Exkurs: Sensitivität, Spezifität, und Odds-Ratio (II) –

Ein Maß des Zusammenhangs, das **unabhängig von der Grundhäufigkeit** eines Merkmals (Basisrate) ist, stellt das Chancenverhältnis (die **Odds-Ratio**) dar.

Die Chance (**odds**), dass ein Ereignis in einer Gruppe G_1 auftritt, ist das Verhältnis der (absoluten oder relativen) Häufigkeit des Auftretens des Ereignisses in dieser Gruppe zur (absoluten oder relativen) Häufigkeit, dass dieses Ereignis in dieser Gruppe nicht auftritt:

$$Odds_{G_1} = \frac{p_{G_1}}{1 - p_{G_1}} = \frac{a}{c} \quad \text{mit } a \text{ und } c = \text{Häufigkeiten in Zellen } a \text{ und } c \text{ einer Kreuztabelle}$$

Genauso kann auch die *Odds* für das Auftreten des Ereignisses in Gruppe G_2 (mit Zellen b und d der Kreuztabelle) berechnet werden. Die **Odds-Ratio** ist das Verhältnis dieser *Odds*:

$$Odds\text{-Ratio} = \frac{p/(1-p)}{q/(1-q)} = \frac{a/c}{b/d} = \frac{a \cdot d}{b \cdot c} \quad \text{mit } p = \text{relative Häufigkeit des Ereignisses in } G_1 \text{ und } q = \text{relative Häufigkeit des Ereignisses in } G_2.$$

Die Odds-Ratio ist das Verhältnis der Chancen des Ereignisses in beiden Gruppen. Technisch wird es auch als **Kreuzproduktverhältnis** der Häufigkeiten in den Zellen der Kreuztabelle bezeichnet.

Beispiel: Die Odds-Ratios für das Ereignis "Krankheit" bei einem positivem Test (s.o.)

Odds-Ratio (Chancenverhältnis):	Odds-Ratio (Chancenverhältnis):
Männer: $OR = (a \cdot d) / (b \cdot c)$	Frauen: $OR = (a \cdot d) / (b \cdot c)$
$= (518 \cdot 183) / (117 \cdot 182) = 4.45$	$= (666 \cdot 61) / (39 \cdot 234) = 4.45$

Die Odds-Ratio ist unabhängig von der Basisrate! Deshalb ist bei gleicher Qualität des Tests die Odds-Ratio bei Männern und Frauen gleich.

Weitere Korrelationskoeffizienten (3)

– Exkurs: Sensitivität, Spezifität, und Odds-Ratio (III) –

Die **Odds-Ratio** sollte **nicht als das Verhältnis von Risiken** interpretiert werden! Korrekt ist die Aussage: "Das Verhältnis von Kranken zu Gesunden ist bei positivem Testergebnis 4.45 mal so hoch wie bei negativem Testergebnis". Nur wenn die **Basisrate klein** ($< .10$) ist, entspricht die Odds-Ratio **näherungsweise dem Risikoverhältnis** (z.B. um wieviel höher als bei negativem Testergebnis ist bei positivem Testergebnis die Wahrscheinlichkeit, krank zu sein?).

Wie die Kovarianz und das χ^2 ist die **Odds-Ratio** als Maß der Enge des Zusammenhangs ungeeignet, da seine Größe nach oben offen ist ($OR < 1$ negativer, $OR = 1$ kein und $OR > 1$ positiver Zusammenhang). Eine Normierung auf den Wertebereich -1 bis $+1$ stellt der Y -Koeffizient von Yules dar:

$$Y = \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1} \quad \text{Für obiges Beispiel: Männer und Frauen: } Y = 0.357$$

bleiben Sensitivität und Spezifität jeweils unverändert, ergeben Korrelationsmaße, die auf der **Kreuzprodukt-Differenz** ($a \cdot d - b \cdot c$) basieren (ϕ -Koeffizient, punkt-biseriale Korrelation = Produkt-Moment-Korrelation), in Abhängigkeit von der Häufigkeit des zu erkennenden Merkmals unterschiedliche Werte.

Das Ergebnis des ϕ -Koeffizienten im o.g. Beispiel ist aber nicht falsch: Man kann bei Männern besser vom Testergebnis auf die Krankheit schließen. Aber nicht weil der Test besser ist, sondern weil die Krankheit bei Männern häufiger auftritt.

Entscheidend für die Wahl des Koeffizienten ist also auch die Fragestellung!

Soll die Qualität des Tests selbst beurteilt werden, muss ein Korrelationskoeffizient gewählt werden, der von der Auftretenswahrscheinlichkeit unabhängig ist. Das sind z.B. (wie Yules Y) auf der Odds-Ratio, also auf dem **Kreuzprodukt-Verhältnis** ($a \cdot d / (b \cdot c)$) basierende Koeffizienten.