5.3 (Empirische) Unabhängigkeit und χ^2

5.3.1 (Empirische) Unabhängigkeit

Durch den Vergleich der bedingten Häufigkeiten mit den Randhäufigkeiten kann man Zusammenhänge beurteilen

Illustration an einem Beispiel: (Aggression und Fahrzeugklasse)

Empirische Unabhängigkeit: Die beiden Komponenten X und Y eines bivariaten Merkmals (X, Y) heißen voneinander *(empirisch) unabhängig*, falls für alle $i = 1, \ldots, k$ und $j = 1, \ldots, m$

$$f(b_j|a_i) = f_{\bullet j} = f(b_j) \tag{5.1}$$

und

$$f(a_i|b_j) = f_{i\bullet} = f(a_i) \tag{5.2}$$

gilt.

Satz:

a) Es genügt, entweder (5.1) oder (5.2) zu überprüfen: Mit einer der beiden Beziehungen gilt auch die andere.

b) X und Y sind genau dann empirisch unabhängig, wenn für alle $i = 1, \ldots k$ und alle $j = 1, \ldots m$ gilt:

$$f_{ij} = f_{i\bullet} \cdot f_{\bullet j} \tag{5.3}$$

c) Gleichung (5.3) ist äquivalent zu

$$h_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n} \tag{5.4}$$

Beweis zu b) :

5.3.2 χ^2-Abstand

Zentrale Idee zur Assoziationsanalyse von Kontingenztafeln:

Als Maß verwendet man den sog. χ^2-*Koeffizienten* / χ^2-*Abstand*. Mit

$$\tilde{h}_{ij} := \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}.$$

definiert man

$$\begin{aligned} \chi^2 &:= \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} \qquad (5.5) \\ &= \sum_{\text{alle Zellen}} \frac{\big(\text{beob. Häufigk.} - \text{unter Unabh. zu erwartende Häufigk.}\big)^2}{\text{unter Unabh. zu erwartende Häufigk.}} \end{aligned}$$

Beispiel:
Zusammenhang zwischen Geschlecht und Arbeitslosigkeit (fiktiv, nach Wagschal, 1999)

Sei Y der Beschäftigungsstatus einer erwerbstätigen Person, X das Geschlecht mit

$$Y = \begin{cases} 1 & \text{beschäftigt} \\ 2 & \text{arbeitslos} \end{cases} \qquad \text{und} \qquad X = \begin{cases} 1 & \text{weiblich} \\ 2 & \text{männlich} \end{cases}$$

Gemeinsame Häufigkeitsverteilung:

X \ Y	1	2	
1	40	25	
2	80	5	

Zur Bestimmung des χ^2-Koeffizienten:

1. Bestimme die Randverteilung.
2. Berechne die unter Unabhängigkeit zu erwartenden Häufigkeiten $\tilde{h}_{ij}$.

Man erhält:

Die Formeln gelten für Kreuztabellen beliebiger Größe. Bei Vierfeldertafeln vereinfachen sich die Tabellen wesentlich, mit der Angabe der Häufigkeit in einer Zelle sind bei gegebenen Randhäufigkeiten auch die Häufigkeiten in den anderen Zellen bestimmt.

Bemerkung: Bei Vierfeldertafeln (2 Zeilen, 2 Spalten) gibt es eine handliche Alternative zur Berechnung von χ^2:

$$\chi^2 = \frac{n \cdot (h_{11}h_{22} - h_{12}h_{21})^2}{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}} \tag{5.6}$$

Veranschaulichung der Formel:

5.3.3 χ^2-basierte Maßzahlen

Bemerkungen zum χ^2-Abstand:

- Unter empirischer Unabhängigkeit gilt per Definition $\chi^2 = 0$. Je stärker χ^2 von 0 abweicht, umso stärker ist ceteris paribus, also unter gleichen sonstigen Größen, der Zusammenhang.
- Der χ^2-Abstand wird die Grundlage bilden für den in Statistik 2 betrachteten χ^2-Test.
- Als Masszahl ist χ^2 hingegen problematisch und nicht direkt interpretierbar, da sein Wert vom Stichprobenumfang n und von der Zeilen- und Spaltenzahl abhängt $\Longrightarrow$ geeignet normieren.
- Es gilt: $\chi^2 \leq n \cdot (\min\{k, m\} - 1)$. Gleichheit gilt genau dann, wenn sich in jeder Spalte bzw. Zeile nur ein von Null verschiedener Eintrag befindet, also z.B. nur auf der Diagonalen von Null verschiedene Einträge aufsetzen. Dies entspräche dann einem perfektem Zusammenhang. Allerdings ist eine solche Extremsituation nicht bei allen Randverteilungen möglich.

χ^2-basierte Zusammenhangsmaße

a) Kontingenzkoeffizient nach Pearson:

$$K := \sqrt{\frac{\chi^2}{n + \chi^2}}. \tag{5.7}$$

b) Korrigierter Kontingenzkoeffizient:

$$K^* := \frac{K}{K_{max}} \tag{5.8}$$

mit

$$K_{\max} := \sqrt{\frac{\min\{k, m\} - 1}{\min\{k, m\}}}$$

c) Kontingenzkoeffizient nach Cramér (Cramérs V):

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min\{k,m\} - 1)}} \tag{5.9}$$

$$= \sqrt{\frac{\chi^2}{\text{maximal möglicher Wert von} \chi^2}}$$

d) Bei der Vierfeldertafel $(k = m = 2)$ gilt

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min\{k,m\} - 1)}} = \sqrt{\frac{\chi^2}{n}}$$

Hierfür ist die Bezeichnung *Phi-Koeffizient* Φ üblich.

Mit (5.6) ergibt sich also

$$\Phi = \left| \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}}} \right|. \tag{5.10}$$

Lässt man die Betragsstriche weg, so erhält man den *signierten Phi-Koeffizienten* oder *Punkt-Korrelationskoeffizienten*

$$\Phi_s = \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}}},$$

der häufig ebenfalls als *Phi-Koeffizient* bezeichnet wird.

Φ_S kann im Prinzip Werte zwischen -1 und 1 annehmen (ohne -1 und 1 immer erreichen zu können (s.u.),).

Vorteil gegenüber Φ: Zusätzlich ist die „Richtung“ des Zusammenhangs erkennbar:

$\Phi_s > 0$

und

$\Phi_s < 0$

Bemerkungen

- K, K^*, V und Φ nehmen Werte zwischen 0 und 1 an, wohingegen χ^2 beliebig grosse positive Werte annehmen kann.
- Aufgrund ihrer Unabhängigkeit von n sind K, K^*, V und Φ prinzipiell zum Vergleich verschiedener Tabellen gleicher Größe geeignet; Φ natürlich nur bei Vierfeldertafeln, wegen ihrer Unabhängigkeit von k und m sind K^* und V auch zum Vergleich von Tabellen mit unterschiedlicher Zeilen und Spaltenzahl geeignet.
- Allerdings kann – bei gegebener Randverteilung – der Wert 1 nicht immer erreicht werden. Im Beispiel können bei insgesamt nur 30 Arbeitslosen nicht alle 80 Männer oder alle 65 Frauen arbeitslos sein.
- Es kann deshalb aussagekräftiger sein, noch zusätzlich auf die für die gegebene Randverteilung maximal mögliche Abhängigkeit zu normieren (s.u.).

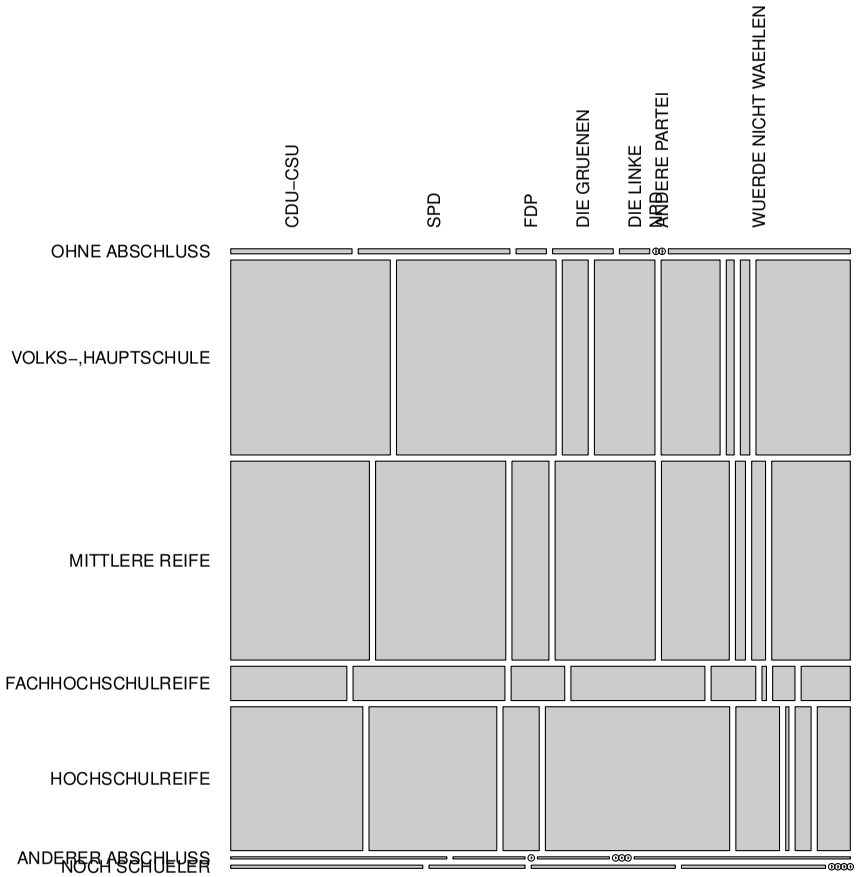
Berechnung im Beispiel: Beschäftigungsstatus und Geschlecht.

Zur Erinnerung: $\chi^2 = 24.435, \quad m = k = 2, \quad n = 150$

besch.		ja	nein	
		1	2	
Frauen	1	40	25	65
Männer	2	80	5	85
		120	30	150

- $K =$
- $K_{max} =$
- $K^* =$
- $V =$
- $\Phi_s =$

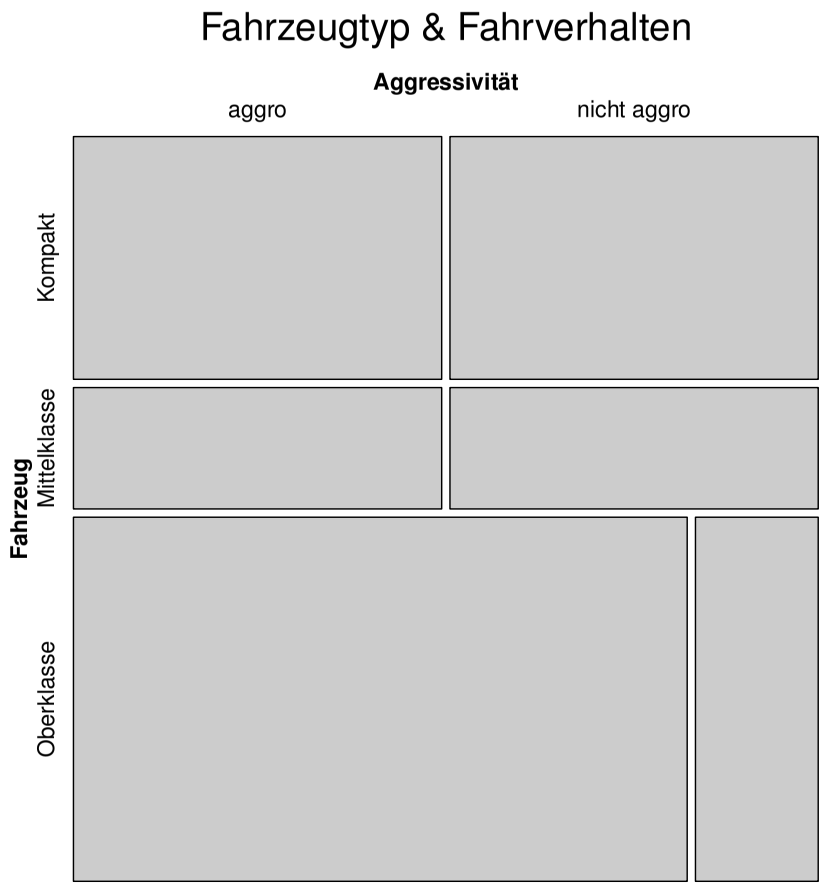
Beispiel 2: Wahlabsicht und Bildungsabschluss (ALLBUS 2010: V327, V747).



Zusammenhangsmaße:

$\chi^2 = 163.71; \qquad K = 0.264; \qquad K^* = 0.284; \qquad V = 0.112$

Beispiel 3: Aggression und Fahrzeugmodell.



Zusammenhangsmaße:

$\chi^2 = 1.5;$ $K = 0.333;$ $K^* = 0.471;$ $V = 0.354$

Korrekturverfahren für Φ (Grundidee nach Wagschal (1999), hier in adaptierter Form: normiere auf den maximal möglichen Wert bei den gegebenen Randverteilungen)

1. Bilde die „strukturtreue $Extremtabelle$“ mit Einträgen h'_{ij}, d.h.

   i. Berechne das Vorzeichen von Φ_s:
   Ist $h_{11} \cdot h_{22} - h_{12} \cdot h_{21} > 0$, so setze $min(h_{12}, h_{21})$ auf 0.
   Ist $h_{11} \cdot h_{22} - h_{12} \cdot h_{21} < 0$, so setze $min(h_{11}, h_{22})$ auf 0.

   ii. Fülle die Tafel entsprechend der Randverteilung auf!

2. Berechne den zugehörigen Phi-Koeffizienten Φ_{extrem}.

3. Berechne den $korrigierten\ (signierten)\ Phi\text{-}Koeffizienten$

$$\Phi_{korr} := \frac{\Phi}{\Phi_{extrem}} \qquad \text{bzw.} \qquad \Phi_{s,korr} := \frac{\Phi_s}{\Phi_{extrem}}.$$

Berechnung im Beispiel:

X \ Y	1	2	
1	40	25	
2	80	5	

5.4 Weitere Methoden für Vierfeldertafeln

Methoden aus der Medizin, die auch in den Sozialwissenschaften mittlerweile große Bedeutung haben. Typische Fragestellung aus der Medizin:

			Y	
			an Krebs erkrankt	nicht an Krebs erkrankt
			b_1	b_2
X	exponiert: Schadstoffen ausgesetzt	a_1	h_{11}	h_{12}
	nicht exponiert: Schadstoffen nicht ausgesetzt	a_2	h_{21}	h_{22}

In der Medizin bezeichnet man die bedingte relative Häufigkeit $f(b_j|a_i)$ als $Risiko$ für b_j unter Bedingung a_i:

$$R(b_j|a_i) := f(b_j|a_i) = \frac{h_{ij}}{h_{i\bullet}} \qquad i,j = 1,2.$$

In der Epidemiologie wird standardmäßig $R(b_1|a_1)$ betrachtet. Dies ist das Erkrankungsrisiko für Personen, die exponiert waren.

Als Zusammenhangsmaß zwischen X und Y in Vierfelder-Tafeln verwendet man auch das darauf aufbauende $relative\ Risiko$:

5.4.1 Relatives Risiko und Prozentsatzdifferenz

Definition: Für eine Vierfelder-Tafel heißt

$$RR(b_1) := \frac{f(b_1|a_1)}{f(b_1|a_2)} = \frac{h_{11}/h_{1\bullet}}{h_{21}/h_{2\bullet}}$$

relatives Risiko. Es betrachtet das Verhältnis des Erkrankungsrisikos für Personen, die exponiert waren (im Zähler) und für Personen, die nicht exponiert waren (im Nenner).

Eigenschaften:

- $RR(b_1)$ kann Werte zwischen 0 und ∞ annehmen.
- $RR(b_1) = 1$ würde bedeuten:
- $RR(b_1) = 5$ würde bedeuten:
- $RR(b_1) = \frac{1}{5}$ würde bedeuten:

In der Medizin bezieht sich „Risiko“ meist auf negative Ereignisse wie z.B. Erkrankung. Grundsätzlich sind Risiken aber symmetrisch verwendbar, d.h. auch für positive Ereignisse wie z.B. Beschäftigung:

beschäftigt	ja	nein	
	1	2	
Frau 1	40	25	65
Mann 2	80	5	85
	120	30	150

Gemessen wird jetzt das „Risiko“ (bzw. besser die Tendenz), beschäftigt zu sein, wenn man dem (vermuteten) Nachteilsfaktor weiblich zu sein, ausgesetzt ist.

Definition:
Die Größe

$$d\%(b_j) := \big(f(b_j|a_1) - f(b_j|a_2)\big) \cdot 100, \qquad j = 1, 2$$

heißt $Prozentsatzdifferenz$ für b_j.

Eigenschaften:

- $d\%(b_1)$ ist z.B. die Differenz aus den Erkrankungsrisiken für Personen, die exponiert waren, und für Personen, die nicht exponiert waren.
- $d\%(b_j)$ kann Werte zwischen -100 und 100 annehmen.
- $d\%(b_1) = 0$ würde bedeuten:
- $d\%(b_1) = 10$ würde bedeuten:
- $d\%(b_1) = -10$ würde bedeuten:

Beispiel: Beschäftigung von Männern und Frauen

beschäftigt	ja	nein	
	1	2	
Frau 1	40	25	65
Mann 2	80	5	85
	120	30	150

Offensichtlich gilt bei zwei Ausprägungen

$$\begin{aligned} d\%(b_1) &= (f(b_1|a_1) - f(b_1|a_2)) = \\ &= (1 - f(b_2|a_1)) - (1 - f(b_2|a_2)) \\ &= -(f(b_2|a_1)) - f(b_2|a_1) = \\ &= -d\%(b_2) \end{aligned}$$

Bemerkungen:

- Den in diesem Abschnitt betrachteten Maßzahlen ist gemein, dass – im Gegensatz zu den χ^2-basierten Maßzahlen – das Vertauschen von Zeilen und Spalten die Maßzahl verändert. Das bedeutet für die Praxis: Man muss sich sehr genau überlegen, was man als abhängige und was als unabhängige Variable wählt.

- Man kann die zwei Risiken in einer Vierfelder-Tafel auf zwei Arten vergleichen:
  - durch den Quotienten: sind Zähler und Nenner eines Bruches gleich, hat er den Wert 1 (d.h. die 1 dient als Vergleichswert)

    $\Rightarrow$ der Bruch ist > 1, wenn der Zähler größer ist als der Nenner.

    $\Rightarrow$ der Bruch ist < 1, wenn der Zähler kleiner ist als der Nenner.
  - durch die Differenz: sind die beiden Terme einer Differenz gleich, hat sie den Wert 0 (d.h. die 0 dient als Vergleichswert)

    $\Rightarrow$ die Differenz ist > 0, wenn der erste Term größer ist als der zweite.

$\Rightarrow$ die Differenz ist < 0, wenn der erste Term kleiner ist als der zweite.

- Bei kleinen Risiken ist die Prozentsatzdifferenz nicht sensitiv, z.B.:
  - $f(b_1|a_1) = 0.42,\ f(b_1|a_2) = 0.41$

    $RR(b_1) = 1.02$

    $d\%(b_1) = 1$
  - $f(b_1|a_1) = 0.02,\ f(b_1|a_2) = 0.01$

    $RR(b_1) = 2.0$

    $d\%(b_1) = 1$

  In solchen Fällen muss man besonders stark inhaltlich abwägen, ob der Quotient oder die Differenz aussagekräftiger ist.

5.4.2 Odds Ratio

Definition: Die Größe

$$O(b_1|a_i) := \frac{R(b_1|a_i)}{1 - R(b_1|a_i)} \qquad i = 1, 2$$

heißt $Odds$ oder $Chance$ von b_1 unter der Bedingung a_i.

Eigenschaften:

- Die $Odds$ für exponierte Personen sind das Verhältnis des Risikos, krank zu werden (im Zähler), zum Risiko, nicht krank zu werden, also 1 − dem Risiko krank zu werden (im Nenner).
- Es gilt:

$$\begin{aligned} O(b_1|a_i) &= \frac{f(b_1|a_i)}{1 - f(b_1|a_i)} = \frac{f(b_1|a_i)}{f(b_2|a_i)} \\ &= \frac{h_{i1}/h_{i\bullet}}{h_{i2}/h_{i\bullet}} = \frac{h_{i1}}{h_{i2}} \end{aligned}$$

- Interpretation: Odds $O(b_1|a_1) = 3$ bedeuten, dass exponierte Personen $3\times$ häufiger krank werden, als dass sie gesund bleiben.
- Interpretation als Wettchance: Odds $O(b_1|a_1) = 3$ bedeuten "ich wäre bereit im Verhältnis $3:1$ zu wetten, dass eine exponierte Person krank wird".

Beispiel: Beschäftigung von Männern und Frauen

beschäftigt	ja	nein	
	1	2	
Frau 1	40	25	65
Mann 2	80	5	85
	120	30	150

$O(\text{beschäftigt}|\text{weiblich})$

$O(\text{beschäftigt}|\text{männlich})$

Eine Chance für sich sagt noch nichts über den Zusammenhang zwischen X und Y aus. Wenn es unter den Exponierten halb so viele Kranke wie Gesunde gibt, so kann dies gut oder schlecht sein. Dies hängt von den Odds bei den Nichtexponierten ab. Daher verwendet man als Zusammenhangsmaß zwischen X und Y die relativen Odds, die als $Odds\ Ratio$ bezeichnet werden.

Definition:
Die Größe

$$OR(b_1) := \frac{O(b_1|a_1)}{O(b_1|a_2)}$$

heißt $Odds\ Ratio$ und vergleicht die Odds von exponierten Personen (im Zähler) und nicht exponierten Personen (im Nenner).

Eigenschaften:

- OR kann Werte zwischen 0 und ∞ annehmen.
- $OR = 1$ würde bedeuten:
- $OR = 5$ würde bedeuten:
- $OR = \frac{1}{5}$ würde bedeuten:
- Um die Asymmetrie des Wertebereichs, $[0; 1)$ bei gegenläufigem Zusammenhang und $(1, \infty]$ bei gleichgerichtetem Zusammenhang, zu umgehen, wird gelegentlich auch $\ln OR$ betrachtet. Sein Wertebereich ist $(-\infty, \infty)$, wobei nun der Wert 0 auf keinen Zusammenhang hinweist, aber dennoch ist der Abstand zu 0 nicht gleich.

- Der *Odds Ratio* wird auch als *Kreuzproduktverhältnis* bezeichnet, denn es gilt:

$$\begin{aligned} OR(b_1) &:= \frac{O(b_1|a_1)}{O(b_1|a_2)} = \frac{\dfrac{R(b_1|a_1)}{1-R(b_1|a_1)}}{\dfrac{R(b_1|a_2)}{1-R(b_1|a_2)}} = \frac{\dfrac{f(b_1|a_1)}{f(b_2|a_1)}}{\dfrac{f(b_1|a_2)}{f(b_2|a_2)}} \\ &= \frac{\dfrac{h_{11}/h_{1\bullet}}{h_{12}/h_{1\bullet}}}{\dfrac{h_{21}/h_{2\bullet}}{h_{22}/h_{2\bullet}}} = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11}\cdot h_{22}}{h_{21}\cdot h_{12}} \end{aligned}$$

Hieraus erkennt man auch die Parallele zu den früheren Zusammenhangsmaßen Φ und χ^2 für 4-Felder-Tafeln, die ebenfalls auf dem Unterschied in den Produkten der Diagonalelemente $h_{11}\cdot h_{22}$ und der Nebendiagonalelemente $h_{12}\cdot h_{21}$ aufbauen. Für χ^2 gilt

$$\chi^2 = \frac{n\cdot(h_{11}\cdot h_{22} - h_{12}\cdot h_{21})^2}{h_{1\bullet}\cdot h_{2\bullet}\cdot h_{\bullet 1}\cdot h_{\bullet 2}}.$$

An dieser Formel erkennt man, dass die Differenz im Zähler

$$h_{11} \cdot h_{22} - h_{21} \cdot h_{12}$$

groß wird, wenn die Häufigkeiten h_{11} und h_{22} auf der Hauptdiagonalen groß, und die Häufigkeiten h_{12} und h_{21} auf den Nebendiagonalen klein sind. Im umgekehrten Fall wird die Differenz klein („stark negativ“).

Durch das Quadrieren des Zählers in der Formel für χ^2 (bzw. durch den Übergang zum Betrag in der Formel für Φ) spielt die Richtung aber keine Rolle mehr, und χ^2 und Φ werden insgesamt groß, wenn

$$h_{11} \cdot h_{22} \gg h_{12} \cdot h_{21}$$

oder

$$h_{11} \cdot h_{22} \ll h_{12} \cdot h_{21}$$

gilt, d.h wenn eine Diagonalstruktur vorliegt, die auf einen Zusammenhang zwischen den Merkmalen Y und X hinweist. („$\ll$“: sehr viel kleiner bzw. größer)

Im OR werden dieselben Häufigkeiten nicht in einer Differenz, sondern in einem Bruch verwendet. Deshalb ist hier nicht von Interesse, ob der Koeffizient von 0 abweicht, wie bei den auf der Differenz aufbauenden Maßzahlen χ^2 und Φ, sondern es interessiert, ob der OR von 1 abweicht.

Beispiel: Beschäftigung von Männern und Frauen

beschäftigt	ja	nein	
	1	2	
Frau 1	40	25	65
Mann 2	80	5	85
	120	30	150

5.4.3 Yules Q

Definition: Die Größe

$$Q := \frac{h_{11} \cdot h_{22} - h_{12} \cdot h_{21}}{h_{11} \cdot h_{22} + h_{12} \cdot h_{21}}$$

heißt Yules Q.

Bemerkungen

- Q ist ein Spezialfall von γ nach Goodman und Kruskal (vgl. später) und vergleicht diskordante und konkordante Paare.
- Q nimmt Werte zwischen -1 und 1 an und ist 0 bei Unabhängigkeit.
- Ist eine Zelle mit 0 besetzt, so ist $Q = 1$ oder $Q = -1$. Q zeigt also dann bereits eine perfekte Abhängigkeit.

Beispiel: Beschäftigung von Männern und Frauen

beschäftigt	ja	nein	
	1	2	
Frau 1	40	25	65
Mann 2	80	5	85
	120	30	150

$$Q = \frac{h_{11} \cdot h_{22} - h_{12} \cdot h_{21}}{h_{11} \cdot h_{22} + h_{12} \cdot h_{21}}$$

$$=$$