5.3 (Empirische) Unabhängigkeit und χ^2

5.3.1 (Empirische) Unabhängigkeit

Illustration an einem Beispiel: (Aggression und Fahrzeugklasse)

Bedingte Häufigkeiten: $f(b_j|a_i)$: relative Häufigkeit von b_j, „wenn man weiß, dass a_i“.

Vergleiche zum Beispiel

$f(b_1|a_3)$: Anteil der Personen mit Ausprägung b_1 (aggressiver Fahrer) unter allen Personen mit Ausprägung a_3 (fährt Oberklasse-Wagen)

mit

$f(b_1) = f_{\bullet 1}$: Anteil der Personen mit b_1 (generell), also alle aggressiven Fahrer Personen

Gilt

$$f(b_1|a_3) > f_{\bullet 1},$$

so erhöht a_3 die Tendenz von b_1 (erhöhte Aggressivität falls Oberklassewagen), gilt

$$f(b_1|a_3) < f_{\bullet 1},$$

so verringert a_3 die Tendenz von b_1 (geringere Aggressivität falls Oberklassewagen). Wäre hingegen

$$f(b_1|a_3) = f_{\bullet 1},$$

so wäre im Beispiel der Anteil der aggressiven Fahrer unter den Oberklassewägen genauso groß wie in der Grundgesamtheit. Das Fahren eines Oberklassewagens würde also nicht das Vorhandensein von aggressivem Fahrverhalten beeinflussen.

Gilt dies für alle Merkmalskombinationen, so beeinflussen sich die Variablen gegenseitig nicht. Die Merkmale sind voneinander unabhängig.

5.3.2 χ^2-Abstand

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	65
2	80	5	85
	120	30	150

Zur Bestimmung des χ^2-Koeffizienten:

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

„Indifferenztafel“(bei empirischer Unabhängigkeit zu erwartende Kontingenztafel):

	1	2	
1	h_{11}	h_{12}	$h_{1\bullet}$
2	h_{21}	h_{22}	$h_{2\bullet}$
	$h_{\bullet 1}$	$h_{\bullet 2}$	n

$$\tilde{h}_{11} = \frac{h_{1\bullet} \cdot h_{\bullet 1}}{n} = \frac{65 \cdot 120}{150} = 52$$

$$\tilde{h}_{21} = \frac{h_{2\bullet} \cdot h_{\bullet 1}}{n} = \frac{85 \cdot 120}{150} = 68$$

etc.

X \ Y	1	2	
1	52	13	65
2	68	17	85
	120	30	150

Beim Vergleich der beoachteten Häufigkeitsverteilung mit der unter Unabhängigkeit zu erwartenden erkennt man: Es gibt 12 weniger beschäftigte Frauen und weniger arbeitslose Männer als bei denselben Randverteilungen zu erwarten wäre, aber mehr arbeitslose Frauen und mehr beschäftigte Männer. Also hat das Geschlecht einen Einfluss; Männer sind tendenziell eher beschäftigt.

Man erhält:

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(h_{ij}-\tilde{h}_{ij})^2}{\tilde{h}_{ij}} \\
&= \frac{(40-52)^2}{52}+\frac{(25-13)^2}{13}+\frac{(80-68)^2}{68}+\frac{(5-17)^2}{17} \\
&= 2.769+11.077+2.118+8.471 \;=\; 24.435
\end{aligned}
$$

Die besprochene Formel zur Berechnung des χ^2-Koeffizienten gilt für Kreuztabellen beliebiger Größe. Bei Vierfeldertafeln vereinfachen sich die Tabellen wesentlich da ja, mit der Angabe der Häufigkeit in einer Zelle bei gegebenen Randhäufigkeiten auch die Häufigkeiten in den anderen Zellen bestimmt sind.

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

$$\chi^2 = \frac{n \cdot (h_{11}h_{22} - h_{12}h_{21})^2}{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}} \tag{5.25}$$

(Merksatz: Hauptdiagonalenprodukt − Nebendiagonalenprodukt).

Berechnung im Beispiel mit alternativer Formel:

$$\begin{aligned} \chi^2 &= n \cdot \frac{(h_{11}h_{22} - h_{12}h_{21})^2}{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}} \\ &= \\ &= \frac{1800 \cdot 1800}{120 \cdot 30 \cdot 65 \cdot 85} \cdot 150 \;=\; 24.434 \end{aligned}$$

5.3.3 χ^2-basierte Maßzahlen

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

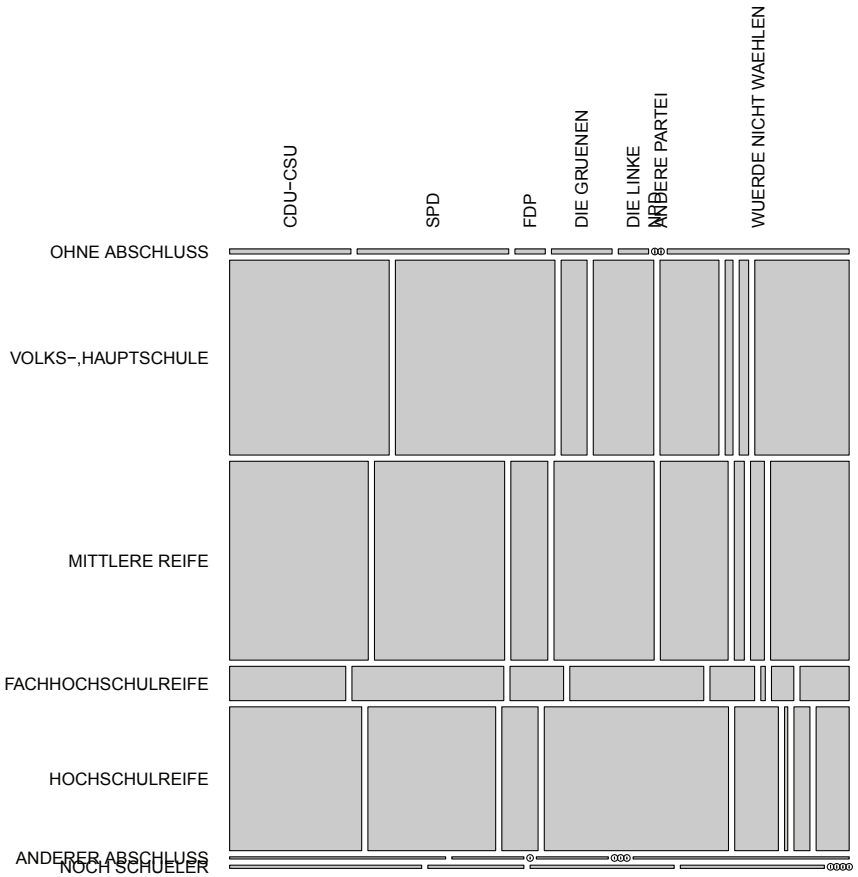
- $K =$
- $K_{max} =$
- $K^* =$
- $V =$

- $\Phi_s =$

$$
\begin{aligned}
K &= \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{24.435}{150+24.435}} = 0.3742 \\
K_{max} &= \sqrt{\frac{\min\{k,m\}-1}{\min\{k,m\}}} = \sqrt{\frac{2-1}{2}} = \sqrt{\frac{1}{2}} \\
K^* &= \frac{K}{K_{max}} = 0.3742 \cdot \sqrt{2} = 0.5292 \\
V &= \Phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{24.435}{150}} = 0.4036 \\
\Phi_s &= \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}}} = \frac{40.5 - 80.25}{\sqrt{120 \cdot 30 \cdot 65 \cdot 35}} = -0.4036
\end{aligned}
$$

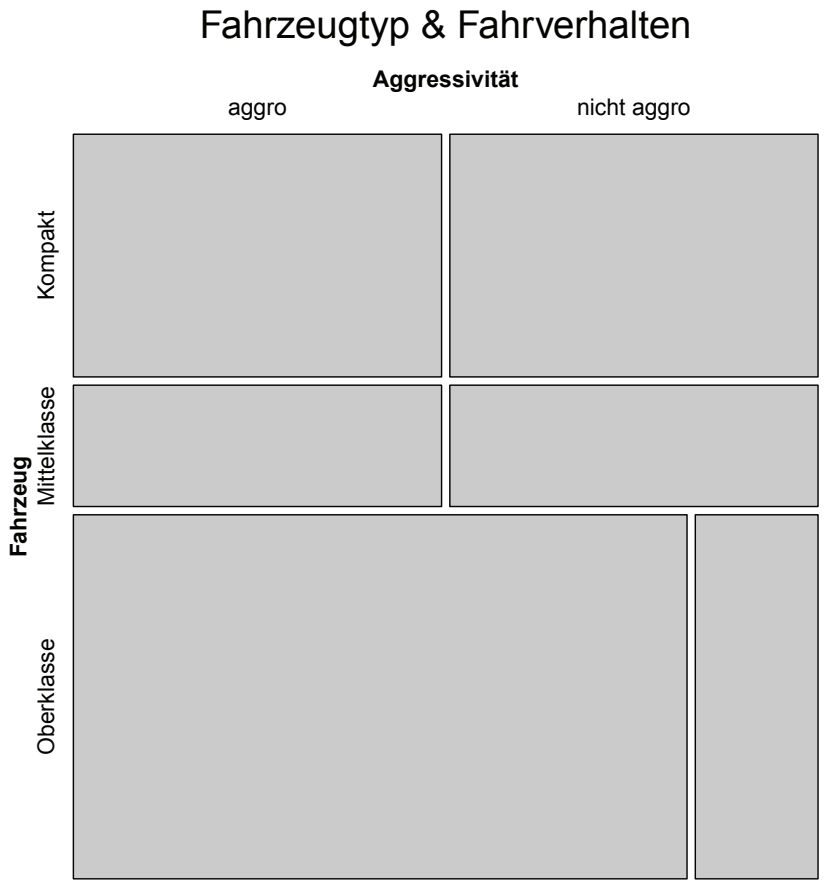
- negatives Vorzeichen: Schwerpunkt auf der Nebendiagonalen: Beschäftigte Männer, arbeitslose Frauen
- Φ =0.4, deutet auf Zusammenhangs mittlerer Stärke hin.

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



Zusammenhangsmaße: $\chi^2 = 163.71; \quad K = 0.264; \quad K^* = 0.284; \quad V = 0.112$

Beispiel 3: Aggression und Fahrzeugtyp.



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

Berechnung im Beispiel:

X \ Y	1	2	
1	40	25	65
2	80	5	85
	120	30	150

$\phi_s = -0.4036$, d.h. setze min(h_{11}, h_{22}) auf 0, also da $h_{11} = 40$ $h_{22} = 0$

X \ Y	1	2	
1			65
2			85
	120	30	150

Jetzt Vierfeldertafel „auffüllen“, so dass Randverteilungen passen.
Extremsituation: Alle Männer beschäftigt.

besch.		ja	nein	
		1	2	
Frauen	1	35	30	65
Männer	2	85	0	85
		120	30	150

Mit der entsprechenden Formel für Φ erhält man

$$\Phi_{extrem} = \left| \frac{h'_{11}h'_{22} - h'_{12}h'_{21}}{\sqrt{h'_{1\bullet}h'_{2\bullet}h'_{\bullet 1}h'_{\bullet 2}}} \right| = \left| \frac{35 \cdot 0 - 30 \cdot 85}{\sqrt{65 \cdot 85 \cdot 120 \cdot 30}} \right| \approx 0.5718$$

und damit

$$\Phi_{korr} = \frac{\Phi}{\Phi_{extrem}} = \frac{0.4036}{0.5718} \approx 0.7059 \quad \text{und} \quad \Phi_{s,korr} \approx -0.7059$$

Relativ zum gegebenen Geschlechterverhältnis und zur Beschäftigungsrelation ergibt sich

ein stärkerer Zusammenhang ($\Phi_{korr} \approx 0.7059$). Am signierten Koeffizienten $\Phi_{s,korr} < 0$ lässt sich auch die Richtung des Zusammenhangs ablesen: kleine Y-Werte gehören eher zu großen X-Werten, also sind Frauen tendenziell stärker von Arbeitslosigkeit betroffen als Männer.

5.4 Weitere Methoden für Vierfeldertafeln

5.4.1 Relatives Risiko und Prozentsatzdifferenz

Beispiel: Beschäftigung von Männern und Frauen

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

$$
\begin{aligned}
d\%(b_1) &= (f(b_1|a_1) - f(b_1|a_2)) \cdot 100 \\
&= \left(\frac{h_{11}}{h_{1\bullet}} - \frac{h_{21}}{h_{2\bullet}}\right) \cdot 100 \\
&= \left(\frac{40}{65} - \frac{80}{85}\right) \cdot 100 = \left(\frac{8}{13} - \frac{16}{17}\right) \cdot 100 = 0.615 - 0.941 \cdot 100 = -32.6
\end{aligned}
$$

Der Beschäftigtenanteil unter den Frauen beträgt 61.5%, der unter den Männer 94.1%. Es ergibt sich eine Prozentsatzdifferenz von 32.6%, die auf einen deutlichen Einfluss des Geschlechts hinweist.

$$\begin{aligned} d\%(b_2) &= (f(b_2|a_1) - f(b_2|a_2)) \cdot 100 = \\ &= \frac{h_{12}}{h_{1\bullet}} - \frac{h_{22}}{h_{2\bullet}} \cdot 100 = 32.6 \end{aligned}$$

Offensichtlich gilt bei zwei Ausprägungen

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

5.4.2 Odds Ratio

Beispiel: Beschäftigung von Männern und Frauen

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

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

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

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

$$O(\text{beschäftigt}|\text{weiblich}) = \frac{h_{11}}{h_{12}} = \frac{40}{25} = \frac{8}{5} = 1.6$$

$$O(\text{beschäftigt}|\text{männlich}) = \frac{h_{21}}{h_{22}} = \frac{80}{5} = 16$$

Unter den Frauen gibt es 1.6 mal so viele Beschäftigte wie Arbeitslose, unter den Männern sind es 16 mal so viele Beschäftigte wie Arbeitslose.

Genau wie ein einzelner Risiko sagt eine Chance für sich 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.

Beispiel: Beschäftigung von Männern und Frauen

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

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

$$
OR(b_1) := \frac{O(b_1|a_1)}{O(b_1|a_2)} = \frac{1}{10}
$$

Frauen haben nur ein Zehntel so hohe Odds für die Beschäftigung im Vergleich zu Männern. Das Verhältnis aus Beschäftigten und Arbeitslosen ist also bei den Frauen um den Faktor 10 geringer als bei den Männern, was für einen starken Zusammenhang

spricht.

5.4.3 Yules Q

Beispiel: Beschäftigung von Männern und Frauen

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

$$Q = \frac{h_{11} \cdot h_{22} - h_{12} \cdot h_{21}}{h_{11} \cdot h_{22} + h_{12} \cdot h_{21}}$$
$$= \frac{40 \cdot 5 - 25 \cdot 80}{40 \cdot 5 + 25 \cdot 80} = -0.818$$

Wieder: starker Zusammenhang in Richtung Nebendiagonale: Männer $\leftrightarrow$ Arbeit, Frauen $\leftrightarrow$ Arbeitslos