5 Assoziationsmessung in Kontingenztafeln

5.1 Multivariate Merkmale

Die Analyse eindimensionaler Merkmale ist nur der allererste Schritt zur Beschreibung der Daten. Meist ist die Analyse von *Zusammenhängen* zwischen Merkmalen von grösserem inhaltlichen Interesse. Beispiele für typische Fragestellung:

- Beeinflusst das Geschlecht das Erwerbseinkommen?
- Gibt es einen Zusammenhang zwischen Schichtzugehörigkeit (als etwas veralteter, dennoch klassischer soziologischer Begriff) und sozialem Engagement?
- Spielt die Stärke der Kirchenbindung eine Rolle bei der Parteienpräferenz?
- Haben Studierende mit guten Mathematikvorkenntnissen bessere Statistiknoten?

Hierzu werden an jeder Einheit $mehrere$ Merkmale erhoben und ihre Ausprägungen auch $gemeinsam$ analysiert (z.B. wird das Geschlecht der i-ten Person mit ihrem Einkommen in Beziehung gesetzt).

Hat man z.B. die Merkmale X, Y, Z und analysiert sie gemeinsam, so nennt man das Paar (X, Y) bzw. das Tripel (X, Y, Z) ein zweidimensionales (bivariates) bzw. dreidimensionales Merkmal (trivariates) Merkmal. Allgemein spricht man von mehrdimensionalen Merkmalen.

$$(X, Y): \ \Omega \longrightarrow (W_X \times W_X)$$

$$\omega \longmapsto (X(\omega), Y(\omega))$$

Achtung:

- Die später folgenden statistischen Verfahren messen die Stärke von Zusammenhängen, aber erlauben keine Aussagen über Kausalität!
- Ob eine kausale Interpretation des Zusammenhangs zulässig ist, hängt davon ab, wie die Daten erhoben wurden.
- Statistische Zusammenhangsmaße können nicht klären:
  - die Richtung des Zusammenhangs (was ist Ursache, was Wirkung?)

    $\Rightarrow$ Längsschnitt-Studie, „cross-lag“ Design
  - ob eine dritte, evtl. unbeobachtete Variable den Zusammenhang verursacht $\Rightarrow$ Experiment

5.2 Kontingenztafeln und bedingte Verteilungen

5.2.1 Gemeinsame Verteilung, Randverteilung, Kontingenztafel

Betrachtet wird ein zweidimensionales Merkmal (X, Y) bestehend aus den diskreten Merkmalen X und Y und die zugehörige Urliste

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Wir wollen ferner annehmen, dass X und Y nur endlich viele (wenige), verschiedene Werte

$$a_1, \ldots, a_i, \ldots, a_k \quad \text{bzw.} \quad b_1, \ldots, b_j, \ldots b_m$$

annehmen können.

Anmerkung: In vielen Büchern (v.a. zur induktiven Statistik) wird statt $a_1, \ldots, a_k$ auch $x_1, \ldots, x_k$ und analog statt $b_1, \ldots, b_m$ auch $y_1, \ldots, y_m$ geschrieben. Bei uns sind aber die (x_i, y_i) Werte der Urliste, x_i also der Wert der i-ten Einheit. Daraus ergibt sich zwar die Doppeldeutigkeit der Laufindizes i und j, wir bleiben jedoch bei dieser Notation um Einheitlichkeit mit Fahrmeir et al. (2009) und Jann (2005) herzustellen.

Beispiel (fiktiv):

$$\text{Fahrzeugmodell } X = \begin{cases} 1, & \text{Kompaktklasse} \\ 2, & \text{Mittelklasse} \\ 3, & \text{Oberklasse} \end{cases}$$

$$\text{aggressives Fahrverhalten } Y = \begin{cases} 1, & \text{ja} \\ 2, & \text{nein} \end{cases}$$

Typische Urliste des zweidimensionalen Merkmals (X, Y):

$$(3,1), (2,2), (2,1), (3,1), (3,2), (3,1), (1,2), (1,1), (1,1), (1,2), (3,1), (3,1)$$

Einheit	X	Y
1	3	1
2	2	2
3	2	1
4	3	1
5	3	2
6	3	1
7	1	2
8	1	1
9	1	1
10	1	2
11	3	1
12	3	1

Achtung:

- Tupel sind – im Gegensatz zu Mengen – $geordnete$ Anordnungen von Zahlen

- Die Tupel sind „gemeinsam indiziert“

Gemeinsame relative und absolute Häufigkeitsverteilung:

$$
\begin{aligned}
h_{ij} &= h(a_i, b_j),\ i = 1, \ldots, k,\ j = 1, \ldots, m, \\
&\quad \text{Anzahl von Beobachtungen mit } x = a_i \text{ und } y = b_j. \\
f_{ij} &= h_{ij}/n = f(a_i, b_j),\ i = 1, \ldots, k,\ j = 1, \ldots, m, \\
&\quad \text{Anteil von Beobachtungen mit } x = a_i \text{ und } y = b_j.
\end{aligned}
$$

Man nennt $(h_{ij}), i = 1, \ldots, k, j = 1, \ldots, m$ und (f_{ij}) die $gemeinsame\ Verteilung$ von (X, Y) in absoluten bzw. relativen Häufigkeiten.

Kontingenztafel / Kontingenztabelle / Kreuztabelle: Darstellung der Häufigkeiten in Form einer $(k \times m)$-dimensionalen Häufigkeitstabelle

	b_1	$\cdots$	b_j	$\cdots$	b_m	
a_1	h_{11}	$\cdots$	h_{1j}	$\cdots$	h_{1m}	$h_{1\bullet}$
a_2	h_{21}	$\cdots$	h_{2j}	$\cdots$	h_{2m}	$h_{2\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_i	h_{i1}	$\cdots$	h_{ij}	$\cdots$	h_{im}	$h_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_k	h_{k1}	$\cdots$	h_{kj}	$\cdots$	h_{km}	$h_{k\bullet}$
	$h_{\bullet 1}$	$\cdots$	$h_{\bullet j}$	$\cdots$	$h_{\bullet m}$	n

Der Punkt steht für Summation über den entsprechenden Index. Es ergeben sich die *Randverteilungen* $h_{i\bullet} = \sum_{j=1}^{m} h_{ij} = h_{i1} + \ldots + h_{im} = h(a_i)$, $i = 1, \ldots, k$, für X und $h_{\bullet j} = \sum_{i=1}^{k} h_{ij} = h_{1j} + \ldots + h_{kj} = h(b_j)$, $j = 1, \ldots, m$, für Y. Es gilt also:

- $h_{i\bullet}$ ist die absolute Häufigkeit von a_i,

- $h_{\bullet j}$ ist die absolute Häufigkeit von b_j.

$$
\begin{aligned}
h_{i\bullet} &= \sum_{j=1}^{m} h_{ij}, \\
h_{\bullet j} &= \sum_{i=1}^{k} h_{ij}
\end{aligned}
$$

Also ist $h_{i\bullet}$ die i-te Zeilensumme, $h_{\bullet j}$ die j-te Spaltensumme (daher der Name Randhäufigkeiten).

Kontingenztafel der relativen Häufigkeitsverteilung:

	b_1	$\cdots$	b_j	$\cdots$	b_m	
a_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1m}	$f_{1\bullet}$
a_2	f_{21}	$\cdots$	f_{2j}	$\cdots$	f_{2m}	$f_{2\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{im}	$f_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_k	f_{k1}	$\cdots$	f_{kj}	$\cdots$	f_{km}	$f_{k\bullet}$
	$f_{\bullet 1}$	$\cdots$	$f_{\bullet j}$	$\cdots$	$f_{\bullet m}$	1

mit der relativen Häufigkeiten $f_{ij} = \dfrac{h_{ij}}{n}$ und den $Randverteilungen$

$$f_{i\bullet} = \frac{h_{i\bullet}}{n} = f_{i1} + \ldots + f_{im} = f(a_i),\ i = 1, \ldots, k, \quad \text{für } X$$

und

$$f_{\bullet j} = \frac{h_{\bullet j}}{n} = f_{1j} + \ldots + f_{kj} = f(b_j),\ j = 1, \ldots, m, \quad \text{für } Y.$$

Beispiel: Aggressives Fahren und Fahrzeugmodell (fiktiv)

Beachte: Aus der gemeinsamen Verteilung kann man die Randverteilungen berechnen (aber nicht umgekehrt, s.u.).

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

	CDU-CSU	SPD	FDP	GRÜNE	LINKE	NPD	ANDERE	NICHTW.	
OHNE ABSCHLUSS	4	5	1	2	1	0	0	6	19
VOLKS-,HAUPTSCHULE	203	203	33	77	75	10	12	120	733
MITTLERE REIFE	180	169	48	130	88	13	18	102	748
FACHHOCHSCHULREIFE	26	34	12	30	10	1	5	11	129
HOCHSCHULREIFE	124	120	34	173	41	3	15	31	541
ANDERER ABSCHLUSS	3	1	0	1	0	0	0	3	8
NOCH SCHUELER	4	2	3	3	0	0	0	0	12
	544	534	131	416	215	27	50	273	2190

Bemerkung: Ist $k = m = 2$ so spricht man von einer Vierfeldertafel. Dabei 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.

	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

z.B. gegeben h_{11}
$\Rightarrow h_{12} = h_{1\bullet} - h_{11}$ etc.

Unabhängige und abhängige Variable:

Hat man eine Vermutung über die Richtung einer potentiellen Wirkung, so bezeichnet man die Variablen entsprechend als $unabhängige$ (wirkende, erklärende) und $abhängige$ (bewirkte, erklärte) Variable.

In der Statistik ist es üblich, die unabhängige Variable mit X zu bezeichnen und die abhängige Variable mit Y, wie gewohnt ist dann Y eine Funktion von X.

z.B:

	unabhängige (X)		abhängige (Y)
möglicherweise:	Automodell	$\longrightarrow$	aggressives Fahrverhalten
eindeutig:	Geschlecht	$\longrightarrow$	Einkommen
allgemein:	unabhängige	$\longrightarrow$	abhängige Variable

Damit werden die Häufigkeitsverteilungen für feste Werte der unabhängigen Variablen in den Zeilen angegeben.

Vorsicht: In einigen sozialwissenschaftlichen Büchern wird entgegen dieser Konvention die unabhängige Variable in den Spalten und die abhängige in den Zeilen abgetragen.

5.2.2 Ökologischer Fehlschluss

Es gibt sehr viele gemeinsame Verteilungen, die zu denselben Randhäufigkeiten passen. Im Beispiel oben passen u.a.:

Man sieht also, wie wichtig es zur Feststellung potentieller Zusammenhänge ist, die $gemeinsame$ Verteilung h_{ij} zu kennen, also tatsächlich die Paare (x_i, y_i) zu betrachten.

Der $unzulässige$ Schluss

- von der Randverteilung auf Eigenschaften der gemeinsamen Verteilung,
- also von zwei univariaten Ergebnissen auf ein bivariates,
- von der Kollektiv- auf die Individualebene,

heißt $ökologischer\ Fehlschluss$.

Kommen zwei Eigenschaften (verschiedene Merkmale) häufig vor, heißt dies nicht notwendig, dass sie gemeinsam häufig vorkommen.

5.2.3 Grafische Darstellung der gemeinsamen Verteilung

Verschiedene Darstellungsarten:

- Mosaikplots: Darstellung der gemeinsamen Häufigkeiten h_{ij} als flächentreue Kachelung

- 3D-Säulendiagramm der gemeinsamen Häufigkeiten h_{ij}

- „normale“ Säulendiagramme nach einer Variable aufgespalten, d.h. für jeden Wert a_i von X werden jeweils die Häufigkeiten h_{ij} bzw. f_{ij} aufgetragen.

Mosaikplots

Grafische Darstellung der gemeinsamen Häufigkeiten zweier diskreter Merkmale.

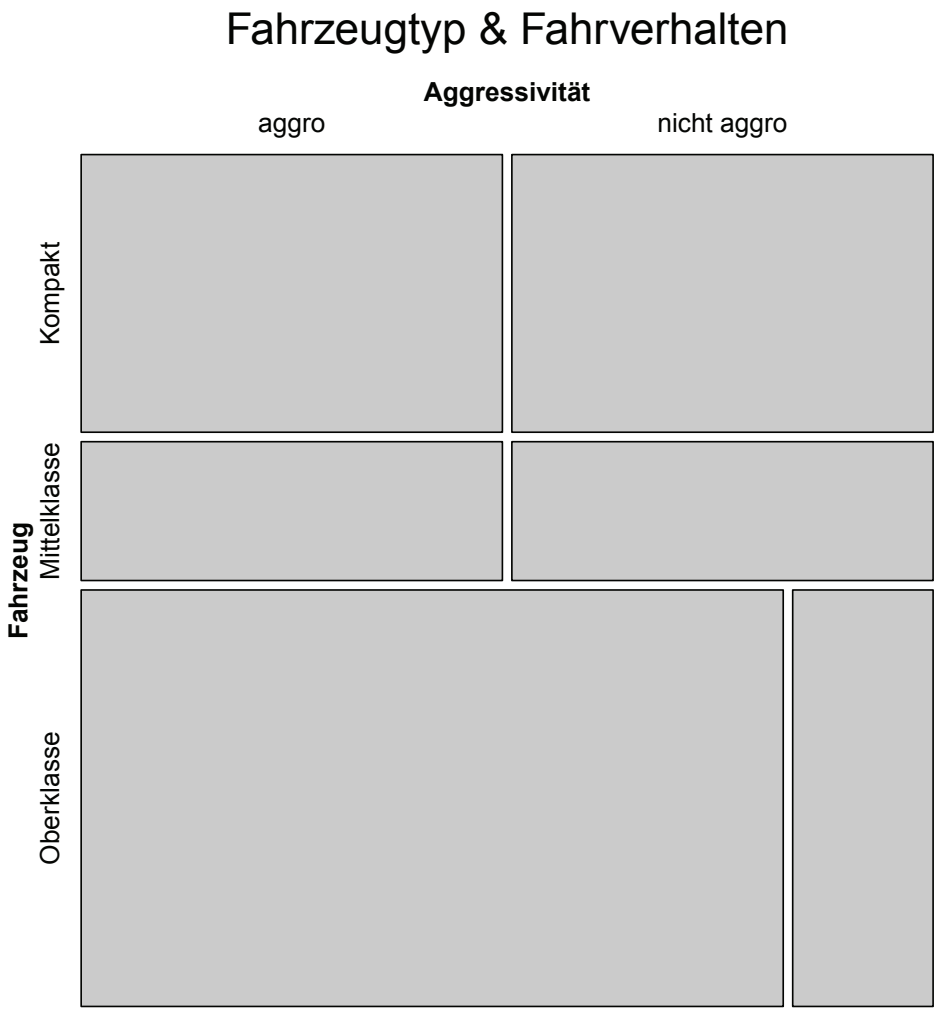
Idee: Teile Quadrat auf in Rechtecke, deren Flächeninhalte f_{ij} entsprechen.

Vorgehen:

1. Teile Einheitsquadrat auf in horizontale Streifen deren Höhe proportional zu $f_{i\bullet}$ ist.

2. Teile horizontale Streifen in Rechtecke deren Breite für jedes i proportional zu f_{ij} ist.

Beispiel: Aggressivität & Fahrverhalten



3D-Säulendiagramm & "Heatmaps" Grafische Darstellung der gemeinsamen Häufigkeiten zweier diskreter Merkmale, auch erweiterbar auf (quasi)-stetige.

Idee: Benutze Merkmalsausprägungen von X, Y als 2-D Koordinatensystem, gemeinsame Häufigkeiten für jede Kombination (a_i, b_j) werden graphisch über Höhe oder über Farbe dargestellt.

Beispiel 1:
Habilitationen nach Geschlecht und Fach (nach Fahrmeir et al., 2009).

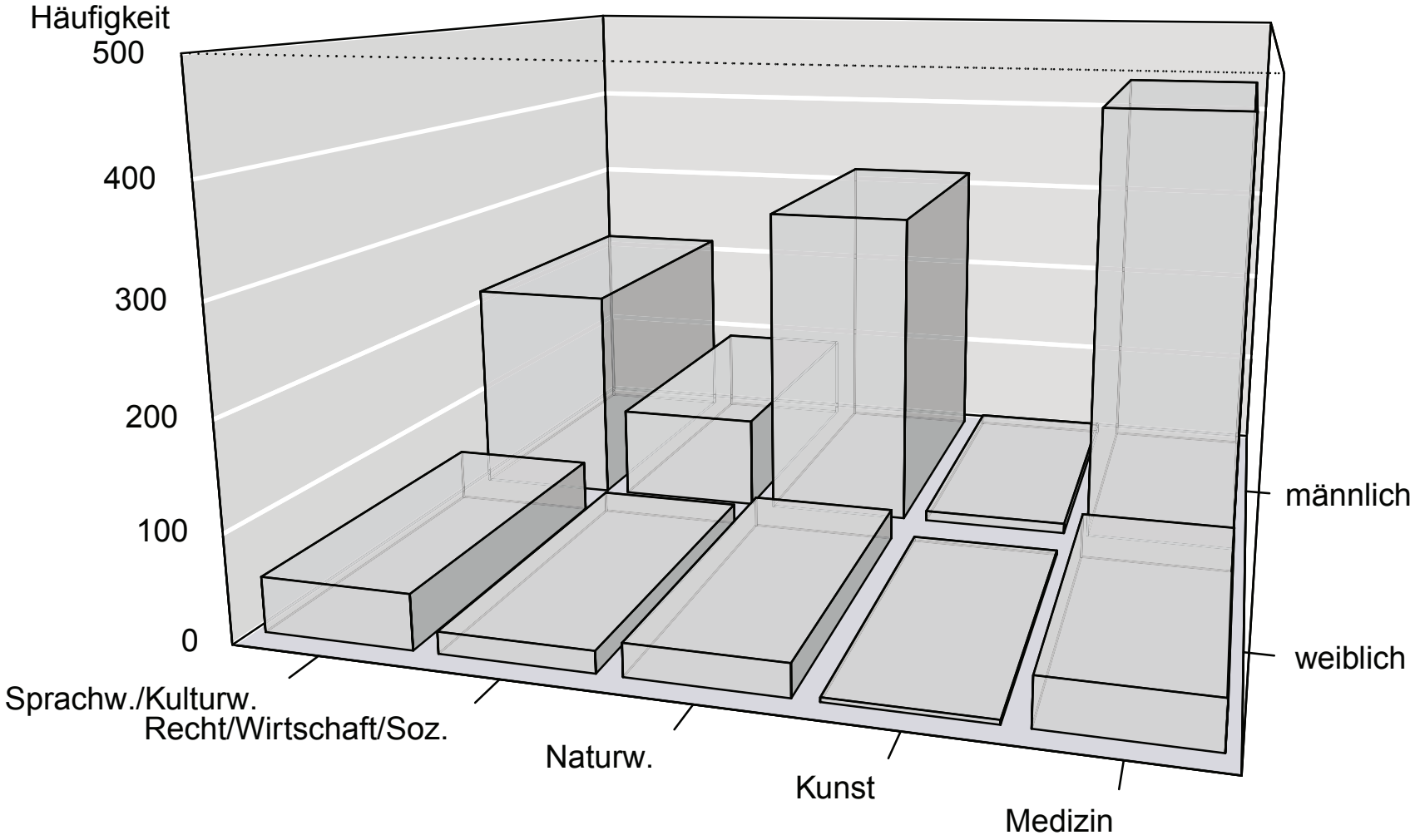
Grundgesamtheit: alle Habilitationen 1993

Geschlecht: X

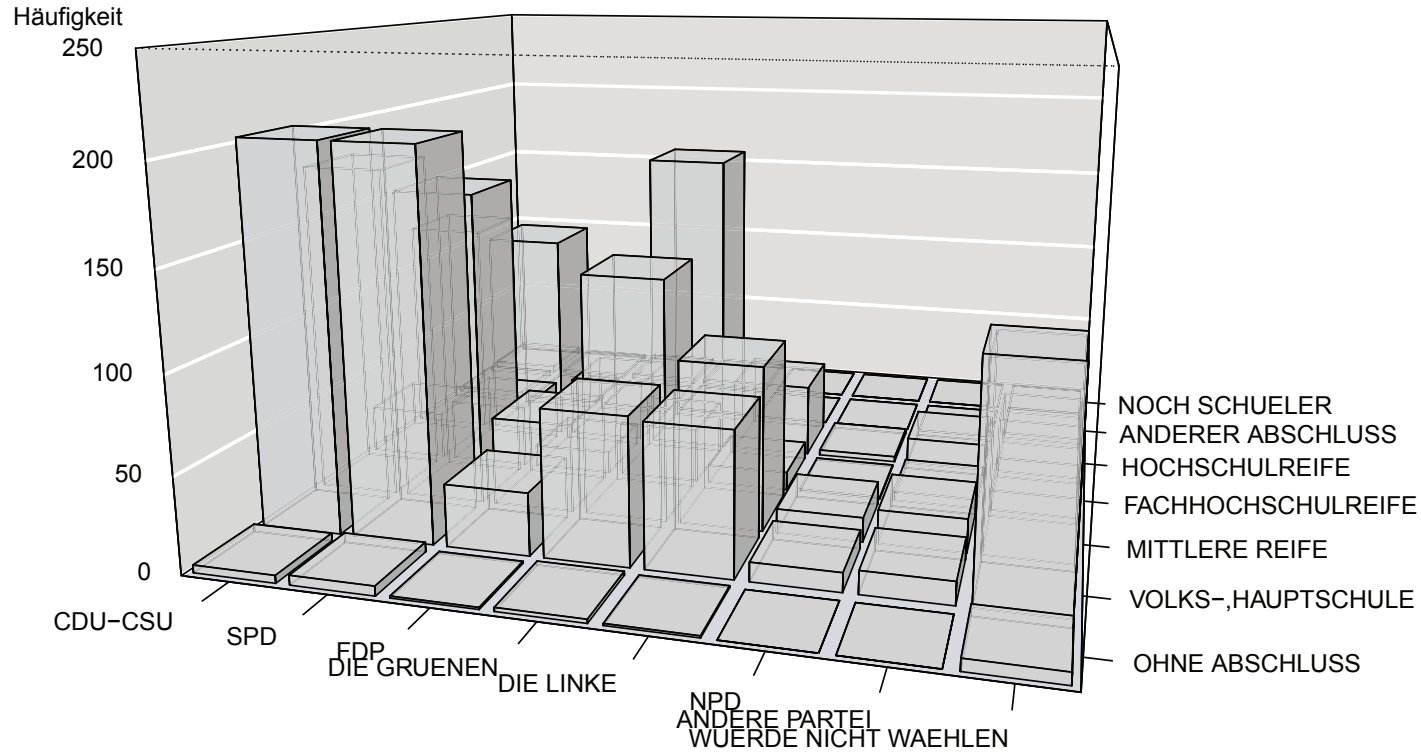
Fächergruppe: Y

		Sprachw. Kulturw.	Rechtsw. Wirts., Soz.	Naturw.	Kunst	Medizin	
		1	2	3	4	5	
weiblich	1	51	20	30	4	44	149
männlich	2	216	92	316	10	433	1067
		267	112	346	14	477	1216

3D-Säulendiagramm:

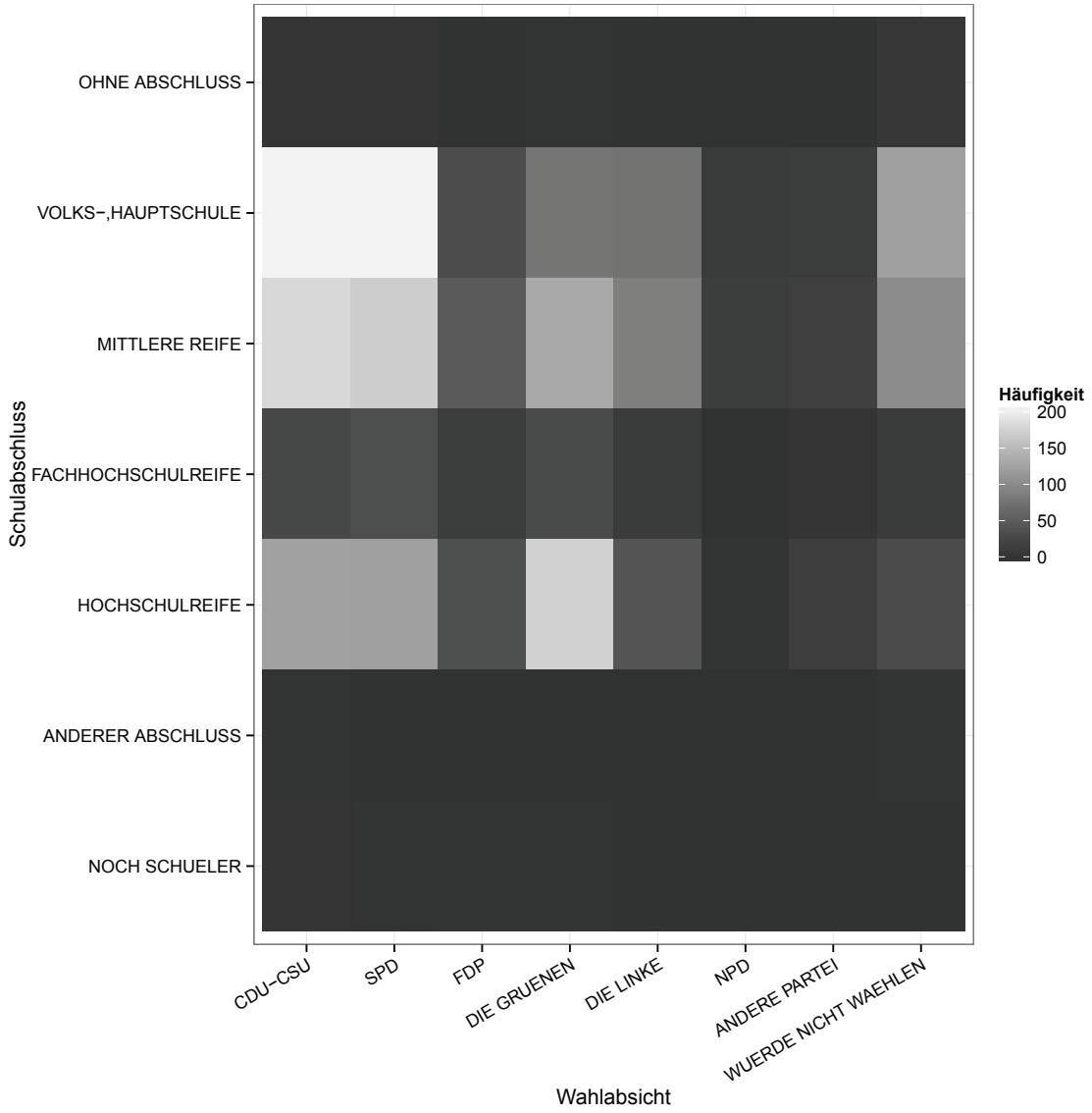


Beispiel 2:
Wahlabsicht und Bildungsabschluss:

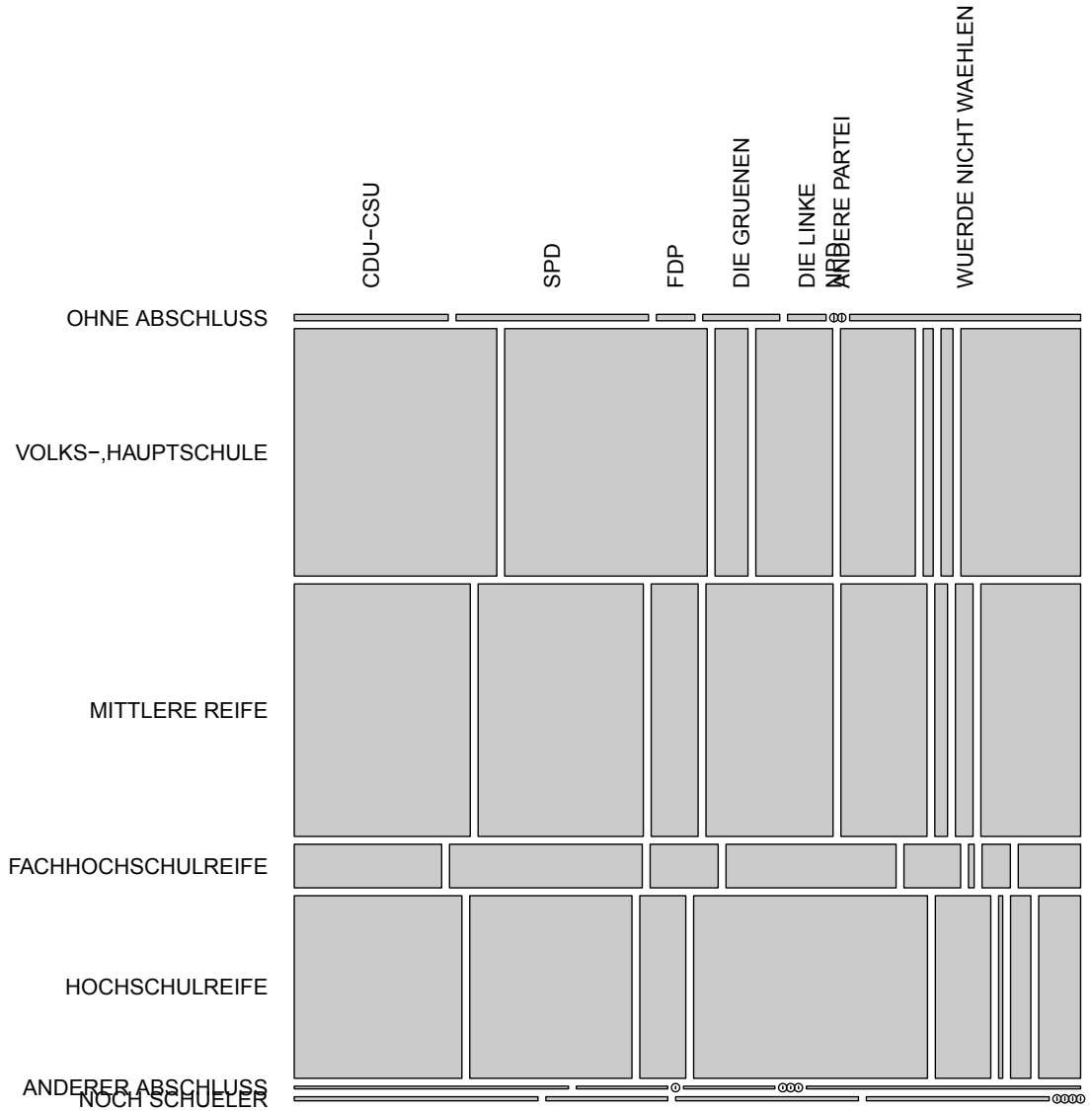


3D-Säulendiagramm hier ungünstig.

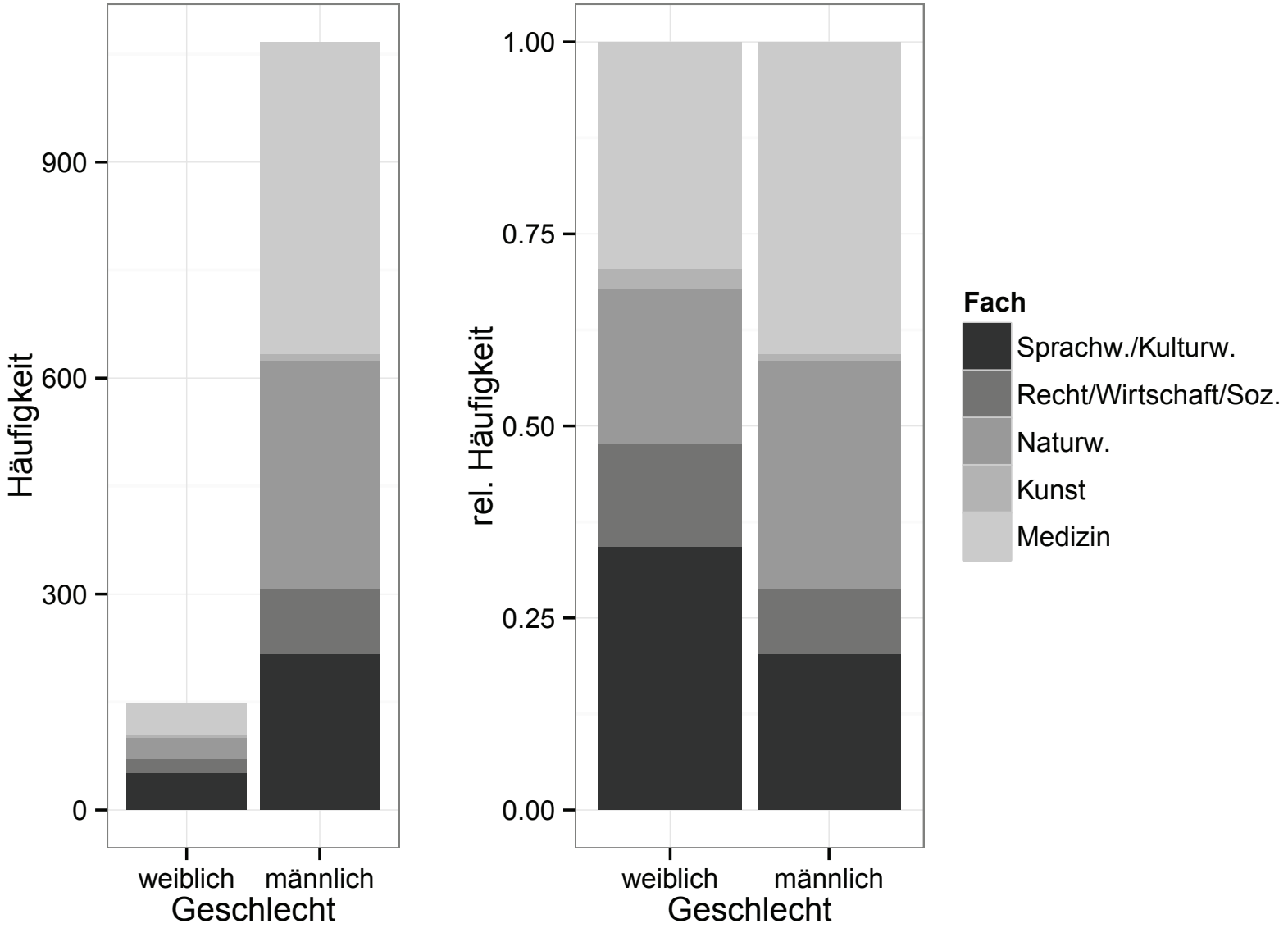
Oft bessere Alternative: Heatmaps (Häufigkeit wird als Farbe kodiert)

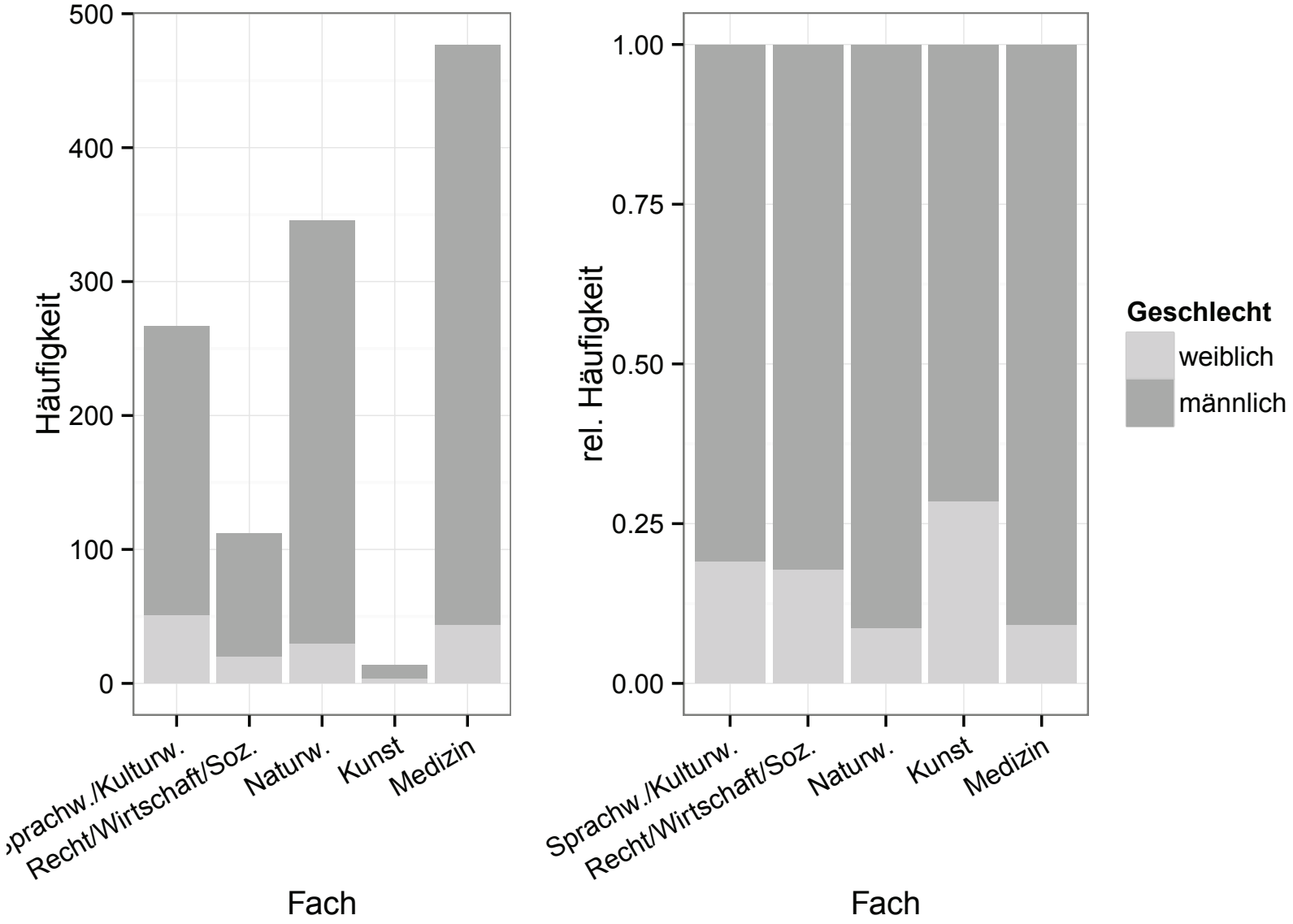


Zum Vergleich der Mosaikplot:



Gestapelte Balkendiagramme & gestapelte skalierte Balkendiagramme




500
400
300
200
100
0
Häufigkeit
Sprachw./Kulturw.
Recht/Wirtschaft/Soz.
Naturw.
Kunst
Medizin
Fach
1.00
0.75
0.50
0.25
0.00
rel. Häufigkeit
Sprachw./Kulturw.
Recht/Wirtschaft/Soz.
Naturw.
Kunst
Medizin
Fach
Geschlecht
weiblich
männlich

5.2.4 Bedingte Häufigkeitsverteilungen
Beispiel:

Habilitationen nach Geschlecht und Fach (nach Fahrmeir et al., 2007).

Grundgesamtheit: alle Habilitationen 1993

Geschlecht: X

Fächergruppe: Y

		Sprachw. Kulturw.	Rechtsw. Wirts., Soz.	Naturw.	Kunst	Medizin	
		1	2	3	4	5	
weiblich	1	51	20	30	4	44	149
männlich	2	216	92	316	10	433	1067
		267	112	346	14	477	1216

Zur Interpretation:

Definition: Seien $h_{i\bullet} > 0$ und $h_{\bullet j} > 0$ für alle i, j. Für jedes $i = 1, \ldots, k$ heißt

$$f_Y(b_1|a_i) := \frac{h_{i1}}{h_{i\bullet}} = \frac{h(a_i, b_1)}{h(a_i)}, \quad \ldots, \quad f_Y(b_m|a_i) := \frac{h_{im}}{h_{i\bullet}} = \frac{h(a_i, b_m)}{h(a_i)}$$

bedingte (relative) Häufigkeitsverteilung von Y unter der *Bedingung* $X = a_i$.

Analog heißt für jedes $j = 1, \ldots, m$

$$f_X(a_1|b_j) := \frac{h_{1j}}{h_{\bullet j}} = \frac{h(a_1, b_j)}{h(b_j)}, \quad \ldots, \quad f_X(a_k|b_j) := \frac{h_{kj}}{h_{\bullet j}} = \frac{h(a_k, b_j)}{h(b_j)}$$

bedingte (relative) Häufigkeitsverteilung von X unter der *Bedingung* $Y = b_j$.

Im Beispiel:

$$f_X(\text{Frau} \mid \text{Habil. in Kunst}) =$$
$$f_X(\text{Frau} \mid \text{Habil. in Naturw.}) =$$

Zu unterscheiden von

$$f_{13} = \frac{h_{13}}{n} = \frac{30}{1216} \approx$$

bzw.

$$f_Y(\text{Habil. in Kunst}|\text{Frau}) =$$

Die Verwechslung von gemeinsamer und bedingter Verteilung bzw. verschiedener bedingter Verteilungen ist eine häufige Fehlerquelle.

Konvention: Bei Vermutung über Richtung des Zusammenhangs betrachtet man vorwiegend die bedingte Verteilung der abhängigen Variablen gegeben die festen Werte der unabhängigen Variable. In diese Richtung geht ja auch die „Prognose“! Man kennt den Wert der unabhängigen Variablen und will Aussagen über die abhängige machen.

Beispiel: Bedingte Verteilung der Fächergruppen gegeben das Geschlecht ($f_Y(b_j|a_i)$ für verschiedene i).

		Sprachw. Kulturw.	Rechtsw. Wirts., Soz.	Naturw.	Kunst	Medizin	
		1	2	3	4	5	
weiblich	1	51	20	30	4	44	149
männlich	2	216	92	316	10	433	1067
		267	112	346	14	477	1216

$Y : b_j$		Sprachw.	Rechtsw.	Naturw.	Kunst	Medizin	
$X : a_i$		Kulturw.	Wirts., Soz.				
		1	2	3	4	5	
weiblich	1						
männlich	2						

$$f(\text{Rechtsw.}|\text{weiblich}) =$$
$$f(\text{Kunst}|\text{männlich}) =$$

Bedingte Verteilung des Geschlechts gegeben die Fachgruppe ($f_X(a_i|b_j)$ für verschiedene j).

b_j / a_i		Sprachw. Kulturw. 1	Rechtsw. Wirts., Soz. 2	Naturw. 3	Kunst 4	Medizin 5
weiblich	1					
männlich	2					

$$f_X(\text{weiblich}|\text{Rechtsw.}) =$$

$$f_X(\text{männlich}|\text{Kunst}) =$$

Nochmals zur Interpretation:

1. $f_X(\text{weiblich}|\text{Medizin}) = .$
2. $f_X(\text{Medizin}|\text{weiblich}) = .$
3. $f_{15} = f(\text{Medizin und weiblich}) = .$

Es liegt jeweils eine andere Grundgesamtheit zu Grunde:

Bedingte Verteilungen werden „automatisch“ durch relative Häufigkeiten ausgedrückt. Für die Berechnung gilt

$$f(a_i|b_j) = \frac{h_{ij}}{h_{\bullet j}} = \frac{\frac{h_{ij}}{n}}{\frac{h_{\bullet j}}{n}} = \frac{f_{ij}}{f_{\bullet j}}$$

und analog

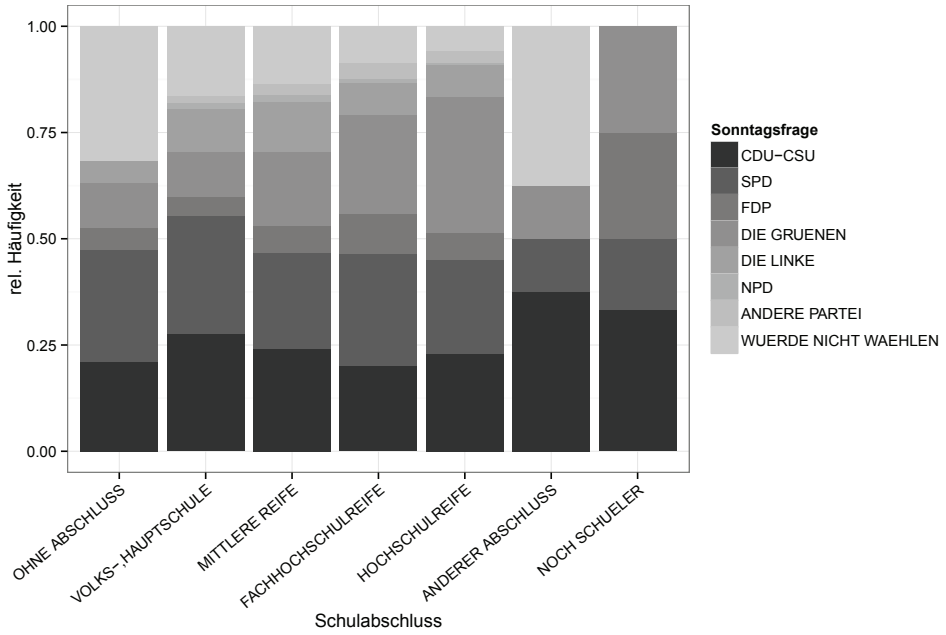
$$f(b_j|a_i) = \frac{h_{ij}}{h_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}.$$

Beispiel: Wahlabsicht und Bildung.

	CDU-CSU	SPD	FDP	GRÜNE	LINKE	NPD	ANDERE	NICHTW.	
OHNE ABSCHLUSS	4	5	1	2	1	0	0	6	19
VOLKS-,HAUPTSCHULE	203	203	33	77	75	10	12	120	733
MITTLERE REIFE	180	169	48	130	88	13	18	102	748
FACHHOCHSCHULREIFE	26	34	12	30	10	1	5	11	129
HOCHSCHULREIFE	124	120	34	173	41	3	15	31	541
ANDERER ABSCHLUSS	3	1	0	1	0	0	0	3	8
NOCH SCHUELER	4	2	3	3	0	0	0	0	12
	544	534	131	416	215	27	50	273	2190

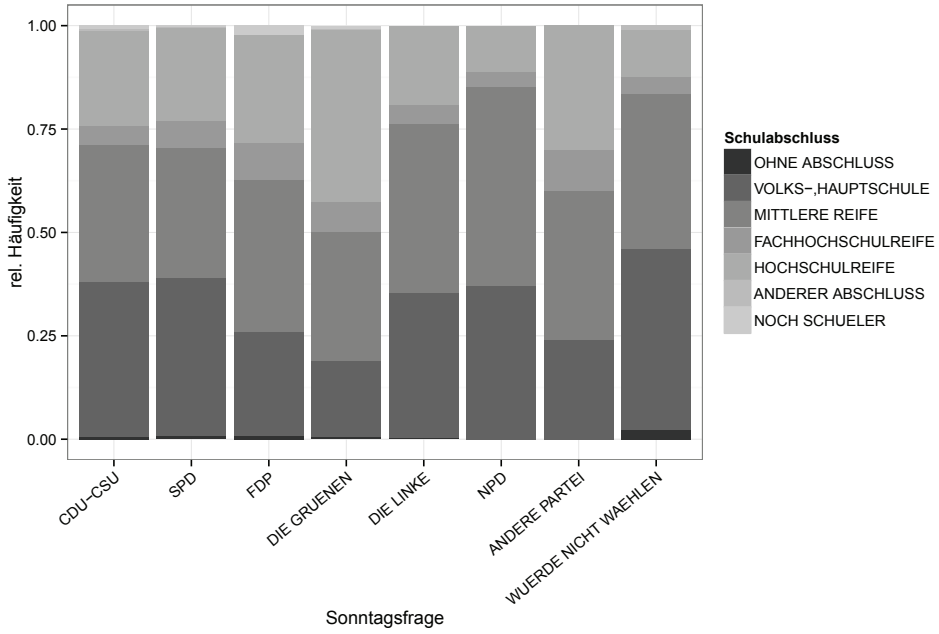
Bedingt auf Bildung:

	CDU-CSU	SPD	FDP	DIE GRUENEN	DIE LINKE	NPD	ANDERE	NICHTW.
OHNE ABSCHLUSS	0.21	0.26	0.05	0.11	0.05	0.00	0.00	0.32
VOLKS-,HAUPTSCHULE	0.28	0.28	0.05	0.11	0.10	0.01	0.02	0.16
MITTLERE REIFE	0.24	0.23	0.06	0.17	0.12	0.02	0.02	0.14
FACHHOCHSCHULREIFE	0.20	0.26	0.09	0.23	0.08	0.01	0.04	0.09
HOCHSCHULREIFE	0.23	0.22	0.06	0.32	0.08	0.01	0.03	0.06
ANDERER ABSCHLUSS	0.38	0.12	0.00	0.12	0.00	0.00	0.00	0.38
NOCH SCHUELER	0.33	0.17	0.25	0.25	0.00	0.00	0.00	0.00
Gesamt	0.25	0.24	0.06	0.19	0.10	0.01	0.02	0.12


1.00
0.75
0.50
0.25
0.00
rel. Häufigkeit
OHNE ABSCHLUSS
VOLKS−,HAUPTSCHULE
MITTLERE REIFE
FACHHOCHSCHULREIFE
HOCHSCHULREIFE
ANDERER ABSCHLUSS
NOCH SCHUELER
Schulabschluss
Sonntagsfrage
CDU−CSU
SPD
FDP
DIE GRUENEN
DIE LINKE
NPD
ANDERE PARTEI
WUERDE NICHT WAEHLEN

Bedingt auf Wahlabsicht:

	CDU-CSU	SPD	FDP	DIE GRUENEN	DIE LINKE	NPD	ANDERE	NICHTW.	Gesamt
OHNE ABSCHLUSS	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.02	0.01
VOLKS-,HAUPTSCHULE	0.37	0.38	0.25	0.19	0.35	0.37	0.24	0.44	0.33
MITTLERE REIFE	0.33	0.32	0.37	0.31	0.41	0.48	0.36	0.37	0.34
FACHHOCHSCHULREIFE	0.05	0.06	0.09	0.07	0.05	0.04	0.10	0.04	0.06
HOCHSCHULREIFE	0.23	0.22	0.26	0.42	0.19	0.11	0.30	0.11	0.25
ANDERER ABSCHLUSS	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
NOCH SCHUELER	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.01


1.00
0.75
0.50
0.25
0.00
rel. Häufigkeit
CDU~CSU
SPD
FDP
DIE GRUENEN
DIE LINKE
NPD
ANDERE PARTEI
WUERDE NICHT WAEHLEN
Sonntagsfrage
Schulabschluss
OHNE ABSCHLUSS
VOLKS~,HAUPTSCHULE
MITTLERE REIFE
FACHHOCHSCHULREIFE
HOCHSCHULREIFE
ANDERER ABSCHLUSS
NOCH SCHUELER

Einschub

(A. Quatember: Unsinn in den Medien - Vom allzu sorglosen Umgang mit Daten)

Man beachte:

- **Es geht nun nicht darum, sich über Fehler anderer Leute lustig zu machen**

- **sondern darum, Sie dafür zu sensibilisieren, dass der Umgang gerade mit bedinten Häufigkeiten sehr viele Fehlerquellen mit sich bringt und große Schwierigkeiten macht.**

$\rightarrow$ **Rüstzeug,**

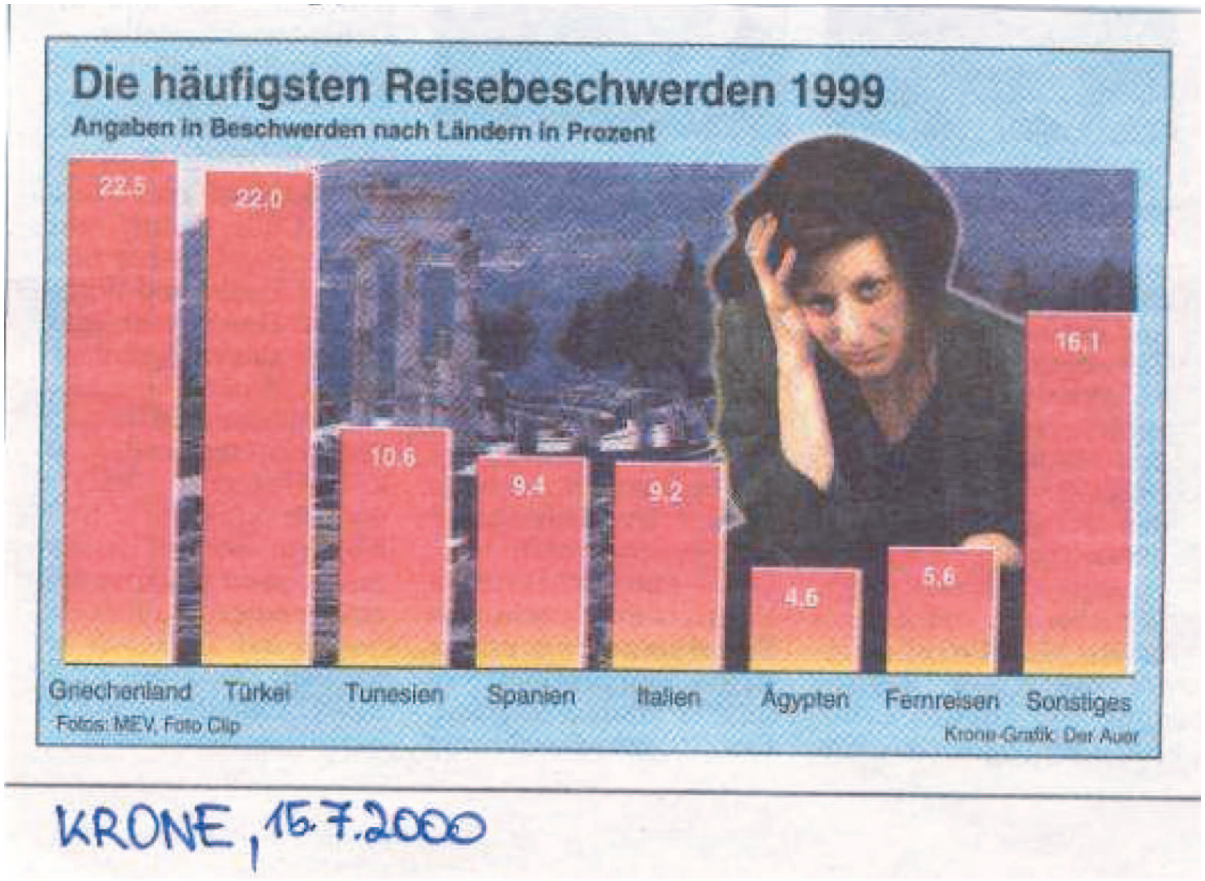
- **damit man selber solche Fehler nicht macht und**

- **Veröffentlichungen kritisch hinterfragen kann.**

Bspl. A. Quatember (Institut für angewandte Statistik, Linz): Unsinn in in den Medien - Vom allzu sorglosen Umgang mit Daten (I):

(https://www.ifas.jku.at/e3456/e3485/e3488/files3493/bedingteverteilung4.pdf?preview=preview)

Man nehme kritisch Stellung zu dem folgenden Zeitungsausschnitt!



Quelle: Kronen-Zeitung, 15.07.2000

Bsp. A. Quatember: Unsinn in in den Medien - Vom allzu sorglosen Umgang mit Daten (II): (https://www.ifas.jku.at/e3456/e3485/e3488/files3489/bedingteverteilung1.pdf?preview=preview)

Wiens Schüler fallen öfter durch

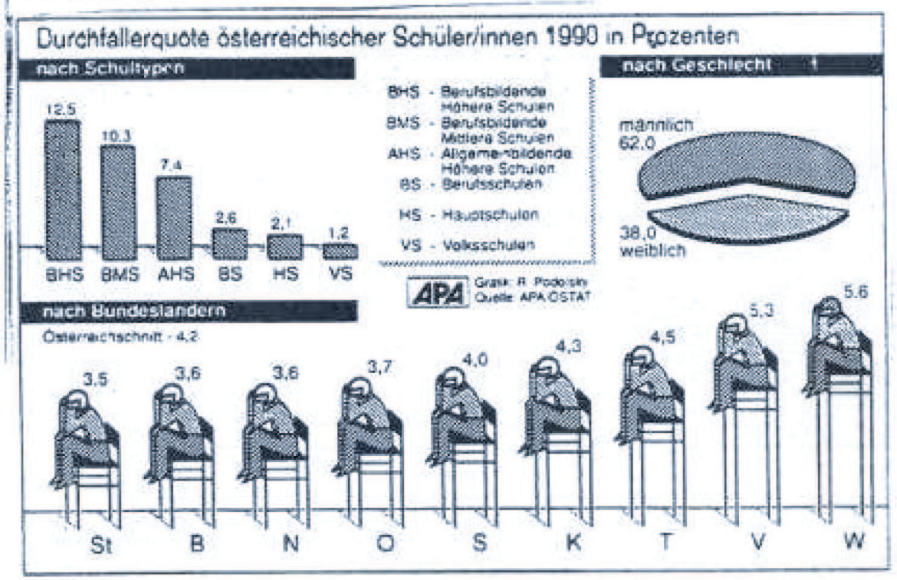
DER STANDARD: 8.5.1992

Mädchen bleiben viel seltener sitzen, sagt das Statistische Zentralamt

Wien – In Wien und Vorarlberg fallen um ein Drittel mehr Schüler durch, als in der Steiermark, Niederösterreich oder im Burgenland. Laut jüngster Erhebung des Österreichischen Statistischen Zentralamtes (ÖSTAT) liegen die „Durchfallerquoten" dieser beiden Länder klar über jenen anderer Bundesländer.

Die Ursachen ortet das ÖSTAT in der unterschiedlichen Leistungsbeurteilung in den Ländern. Im Bundesdurchschnitt dürfen jährlich 4,2 Prozent der Schüler nicht „aufsteigen".

Die Mädchen – oft zahlenmäßig überlegen – stellen nur etwas mehr als ein Drittel der „Sitzenbleiber". Bei den Burschen dagegen erreichen 62 Prozent ihr Klassenziel nicht. Die meisten „Durchfaller" gibt es an Berufsbildenden Höheren Schulen: Nicht einmal jeder Zehnte schafft den Aufstieg. In Höheren Technischen Lehranstalten bleiben mehr als 15 Prozent „sitzen". Die Allgemeinbildenden Höheren Schulen verzeichnen bundesweit 7,4 Prozent „Sitzenbleiber". (APA)



Quelle: Der Standard, 8.05.1992

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 $j = 1, \ldots, m$

$$f(b_j|a_i) = f_{\bullet j} = f(b_j) \tag{5.13}$$

und

$$f(a_i|b_j) = f_{i\bullet} = f(a_i) \tag{5.14}$$

gilt.

Satz:

a) Es genügt, entweder (5.13) oder (5.14) 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:

$$f_{ij} = f_{i\bullet} \cdot f_{\bullet j} \tag{5.15}$$

c) Gleichung (5.15) ist äquivalent zu

$$h_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n} \tag{5.16}$$

Beweis:) Die letzte Gleichung ist auch äquivalent zu $\frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$, also zu $f(b_j|a_i) = f_{\bullet j}$, was a beweist.

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

$$\begin{aligned} \chi^2 &:= \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} \qquad (5.17) \\ &= \sum_{\text{alle Zellen}} \frac{\big(\text{beob. Häufigk.} - \text{unter Unabh. zu erwartende Häufigk.}\big)^2}{\text{unter Unabh. zu erwartende Häufigk.}} \end{aligned}$$

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 Formel 5.17 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:

$$\chi^2 = \frac{n \cdot (h_{11}h_{22} - h_{12}h_{21})^2}{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}} \tag{5.18}$$

Veranschaulichung der Formel:

5.3.3 χ^2-basierte Maßzahlen

Bemerkungen zum χ^2-Abstand:

- Unter empirischer Unabhängigkeit gilt definitionsgemäß $\chi^2 = 0$. 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:

$$K := \sqrt{\frac{\chi^2}{n + \chi^2}}. \tag{5.19}$$

b) Korrigierter Kontingenzkoeffizient:

$$K^* := \frac{K}{K_{max}} \tag{5.20}$$

mit

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

c) Kontingenzkoeffizient nach Cramér (Cramérs V):

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min\{k,m\} - 1)}} \tag{5.21}$$

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

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 (5.18) ergibt sich also

$$\Phi = \left| \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}}} \right|. \tag{5.22}$$

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.).

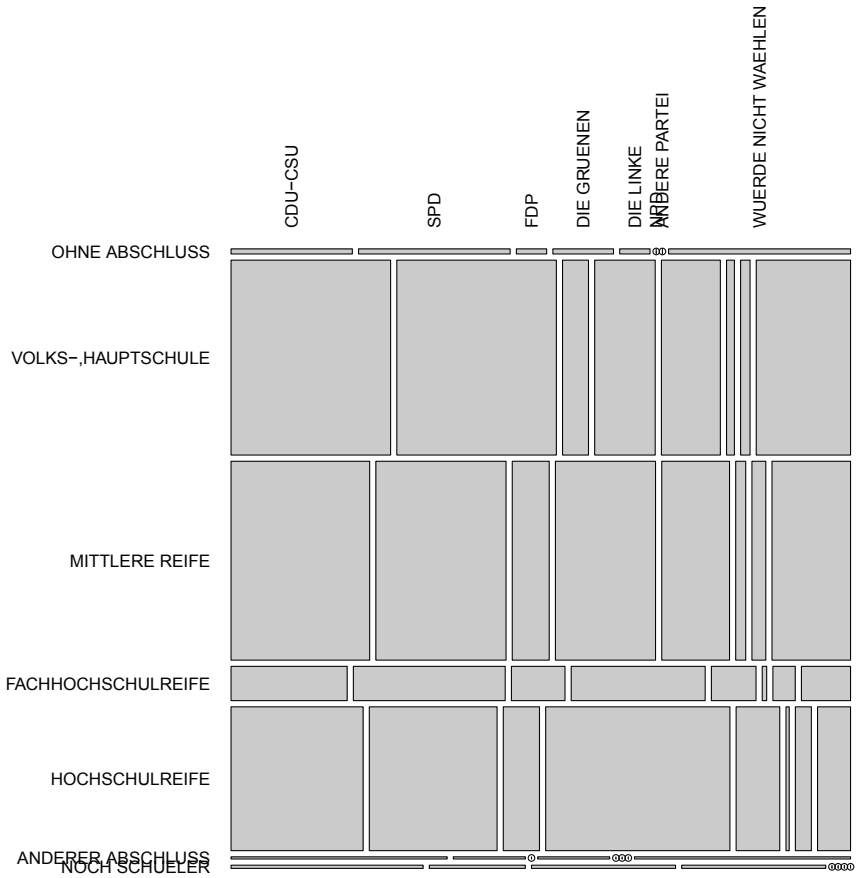
Berechnung im Beispiel: Beschäftigungsstatus und Geschlecht.

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

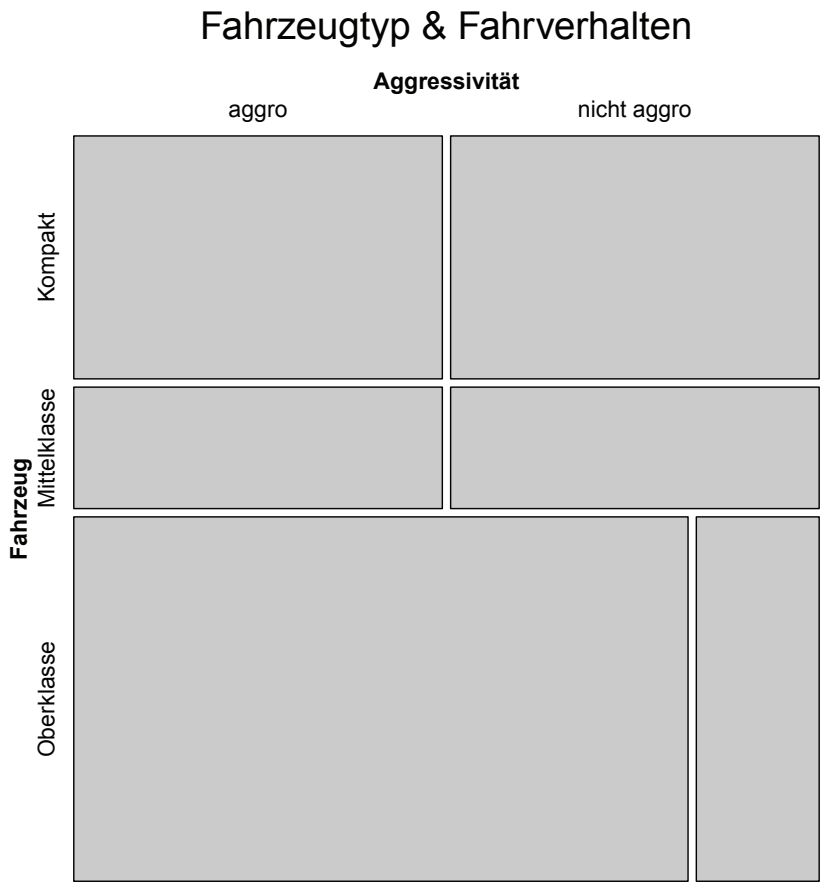
- $K =$
- $K_{max} =$
- $K^* =$
- $V =$
- $\Phi_s =$

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



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

Beispiel 3: Aggression und Fahrzeugtyp.



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

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

1. Denke dir Randverteilungen als fest (gleiches Geschlechterverhältnis, feste Arbeitslosenquote)

2. 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!

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

4. Berechne den *korrigierten (signierten) Phi-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	