Eine und zwei kategorielle Variablen

7.11.2011

Einführung

Kategorielle Variable, Faktor

Eine kategorielle Variable (Faktor) hält fest, zu welcher Kategorie eine Beobachtung gehört.
Falls die Kategorien geordnet werden können: **ordinale**, ansonsten **nominale** kategorielle Variable.

Beispiele

Haarfarbe:	$\{$schwarz, braun, blond, rot$\}$	nominal
Geschlecht:	$\{$m, f$\}$	nominal
Einkommen:	$\{$0–50K, 50–100K, $>$100K$\}$	ordinal

...

In Computer oft via numerische Codes (DB!) abgelegt. Differenzen etc. machen in der Regel keinen Sinn!

Beispiel: Umfrage zum Umweltschutz

Fragen

- Hauptverantwortung für den Umweltschutz:
  bei den Einzelnen / beim Staat / bei beiden
- Beeinträchtigung durch Umweltschadstoffe:
  überhaupt nicht / etwas / ziemlich / sehr beeinträchtigt
- Schulbildung:
  (1) Volks-, Hauptschule ohne Lehrabschluss
  (2) mit Lehrabschluss
  (3) weiterbildende Schule ohne Abitur
  (4) Abitur, Hochschulreife, Fachhochschulreife
  (5) Studium (Universität, Akademie, Fachhochschule)

Zusammenfassung / Darstellung

Zusammenfassung

Bestimme Anzahlen für jede Kategorie.

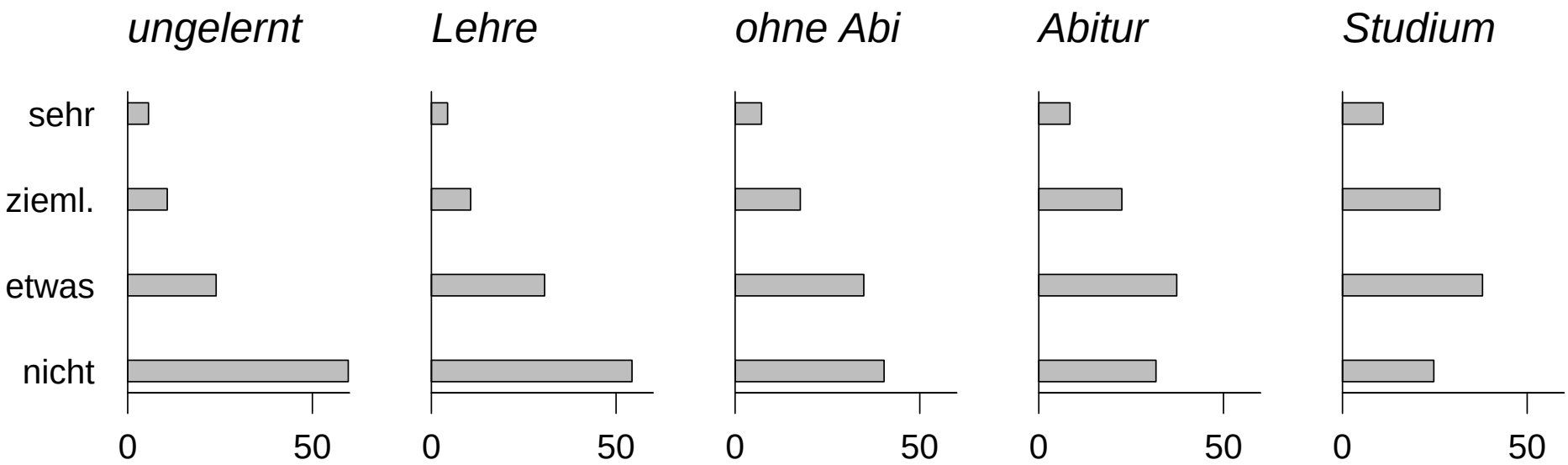
Darstellungen

- Eine Variable $\rightsquigarrow$ Stabdiagramm / Histogramm
- Mehrere Variablen $\rightsquigarrow$ Tabelle (**Kontingenztafel**) oder zeilen/spaltenweise Plots der (relativen) Häufigkeiten.

Beispiel Umfrage: Kontingenztafel

		Beeinträchtigung (B)				
		nicht	etwas	zieml.	sehr	Summe
Schule (A)	ungelernt	212	85	38	20	355
	Lehrabschl.	434	245	85	35	799
	ohne Abi.	169	146	74	30	419
	Abitur	79	93	56	21	249
	Studium	45	69	48	20	182
	Summe	939	638	301	126	2004

Beispiel Umfrage: Relative Häufigkeiten pro Schulabschluss



Häufigkeitsdaten vs. Zähldaten

Häufigkeitsdaten (Frequency Data)

- Zusammenfassung ursprünglicher Beobachtungen von **kategoriellen Variablen**.
- Ursprüngliche Beobachtungen meistens als stochastisch unabhängig vorausgesetzt!
- Fragestellungen betreffen die ursprünglichen Variablen.

Zähldaten (Count Data)

- Ursprüngliche Beobachtungen sind **Anzahlen**.
- Diese Anzahlen können **irgendwie** zustande kommen.
- Bsp: Zähle Anzahl Wildtiere auf Feld bei verschiedenen Witterungsbedingungen.

Wir betrachten heute Frequency Data!

Fragestellungen

- Regressionsansatz
  - **Zielgrösse** (“Antwortfaktor”) und
  - **Ausgangsvariablen** (erklärende Variablen)

  Einführung neuer Modelle:
  - Logistische Regression
  - Multinomiale Regression
  - Kumulative Logits

  Dies betrachten wir später.

- “**Zusammenhänge**”
  - Variablen “gleichberechtigt”.
  - Betrachtung als multivariate Grösse.

Bisher für kontinuierliche Variablen: Regression, Korrelation

Outline

Heute betrachten wir **Zusammenhänge** zwischen kategoriellen Variablen, d.h. die Variablen werden **"gleichberechtigt"** betrachtet.

Programm

- Modelle für Kontingenztafeln
- Test auf Unabhängigkeit
  - ungepaart: χ^2-Test ($r \times s$), Fisher's Exact Test (2×2)
  - gepaart: McNemar (2×2)
- Stärke der Abhängigkeit: Odds Ratio (2×2)

Modelle für Kontingenztafeln (für 2 kategorielle Variablen)

Notation

		Variable B						
		1	2	3		k		s	$\sum$
Variable A	1	n_{11}	n_{12}	n_{13}	$\dots$	n_{1k}	$\dots$	n_{1s}	n_{1+}
	2	n_{21}	n_{22}	n_{23}	$\dots$	n_{2k}	$\dots$	n_{2s}	n_{2+}
		$\vdots$	$\vdots$			$\vdots$		$\vdots$	$\vdots$
	h	n_{h1}	n_{h2}	$\dots$		n_{hk}	$\dots$	n_{hs}	n_{h+}
		$\vdots$	$\vdots$			$\vdots$		$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\dots$		n_{rk}	$\dots$	n_{rs}	n_{r+}
	$\sum$	n_{+1}	n_{+2}	$\dots$		n_{+k}	$\dots$	n_{+s}	n

Brauchen **Wahrscheinlichkeits-Modell**:

$P(A = h, B = k) =: \pi_{hk}$, $\sum_{h,k} \pi_{hk} = 1$

		Variable B						
		1	2	3		k		s	$\sum$
Variable A	1	π_{11}	π_{12}	π_{13}	$\dots$	π_{1k}	$\dots$	π_{1s}	π_{1+}
	2	π_{21}	π_{22}	π_{23}	$\dots$	π_{2k}	$\dots$	π_{2s}	π_{2+}
		$\vdots$	$\vdots$			$\vdots$		$\vdots$	$\vdots$
	h	π_{h1}	π_{h2}	$\dots$		π_{hk}	$\dots$	π_{hs}	π_{h+}
		$\vdots$	$\vdots$			$\vdots$		$\vdots$	$\vdots$
	r	π_{r1}	π_{r2}	$\dots$		π_{rk}	$\dots$	π_{rs}	π_{r+}
	$\sum$	π_{+1}	π_{+2}	$\dots$		π_{+k}	$\dots$	π_{+s}	1

- **Randverteilungen** von A, B: π_{h+}, π_{+k}.
- **Bedingte Vert.** von B, geg. A: $\pi_{k|h} = P(B = k \mid A = h) = \frac{\pi_{hk}}{\pi_{h+}}$.

Schätzung $\widehat{\pi}_{hk} = N_{hk}/n$. $\quad$ (N_{hk} zufällig!)

		Beeinträchtigung (B)				
		nicht	etwas	zieml.	sehr	Summe
Schule (A)	ungelernt	10.6	4.2	1.9	1.0	17.7
	Lehrabschl.	21.7	12.2	4.2	1.7	39.9
	ohne Abi.	8.4	7.3	3.7	1.5	20.9
	Abitur	3.9	4.6	2.8	1.0	12.4
	Studium	2.2	3.4	2.4	1.0	9.1
	Summe	46.9	31.8	15.0	6.3	100.0

Tabelle: Relative Häufigkeiten in Prozenten im Beispiel der Umfrage zu Umweltschadstoffen.

Verteilungen

π_{hk} sind Wahrscheinlichkeiten für **eine** Beobachtung.

n Beobachtungen $\rightsquigarrow$ Erhalte N_{hk}. Verteilung?

- **Total n gegeben (fix)** (*"multinomial sampling"*)

  Bsp: Befrage 2000 Personen nach ihrem Schulabschluss und nach der Beeinträchtigung durch Umweltschadstoffe.

  Multinomiale Verteilung:

$$[N_{11}, N_{12}, ..., N_{rs}] \sim \mathcal{M}(n, \pi_{11}, \pi_{12}, \ldots, \pi_{rs})$$

$$P(N_{11} = n_{11}, N_{12} = n_{12}, \ldots, N_{rs} = n_{rs}) = \frac{n!}{n_{11}! n_{12}! \cdots n_{rs}!} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \cdots \pi_{rs}^{n_{rs}}$$

- **Randtotale von A fest: Geschichtete Stichprobe** (*"independent multinomial sampling"*)

  Bsp: Befrage pro Schulabschluss 400 Personen nach ihrer Beeinträchtigung durch Umweltschadstoffe.

  $N_{h+} = n_{h+}$, r unabhängige Stichproben

  $$[N_{h1}, N_{h2}, \ldots, N_{hs}] \sim \mathcal{M}(n_{h+}, \pi_{h1}, \pi_{h2}, \ldots, \pi_{hs}),$$

  unabh. für $h = 1, \ldots, r$.

- N **zufällig** *("Poisson sampling")*

  Bsp: Befrage 1 Woche lang Personen nach Schulabschluss und nach ihrer Beeinträchtigung durch Umweltschadstoffe.

  Modell der unabhängigen Poisson-Verteilung

  $N_{hk} \sim \mathcal{P}(\pi_{hk} \cdot \lambda)$, unabhängig für $h = 1, \ldots, r$ und $k = 1, \ldots, s$

$$P(N_{11} = n_{11}, N_{12} = n_{12}, \ldots, N_{rs} = n_{rs}) = \prod_{h,k} \frac{\lambda_{hk}^{n_{hk}}}{n_{hk}!} e^{-\lambda_{hk}}.$$

  Hält man N hier fest, so ist die bedingte Verteilung der N_{hk}, gegeben $N = n$, gerade die multinomiale Verteilung von oben.

  Randtotale $N_{h+} = n_{h+}$ festhalten: unabhängige multinomiale Verteilungen.

Unabhängigkeitstests und Vergleich von Stichproben

Modell für **Unabhängigkeit** (Nullhypothese)

$$\pi_{hk} = P(A = h, B = k) = P(A = h) \cdot P(B = k) = \pi_{h+}\pi_{+k}.$$

Oder: Die bedingten Verteilungen von A gegeben B sind gleich:

$$\pi_{k|h} = \pi_{k|h'} \text{ für alle } h.$$

Anschaulich: Egal in welcher Spalte ich schaue, die bedingten Verteilungen sehen immer gleich aus (illustrativ: Barplots).

D.h. wenn ich B kenne, nützt mir das *nichts*, um etwas über A auszusagen.

Wie können wir dies testen? Betrachte standardisierte Residuen.

$$\text{Residuum}_{hk} = \frac{\text{beobachtet}_{hk} - \widehat{\text{erwartet}}_{hk}}{\sqrt{\widehat{\text{erwartet}}_{hk}}},$$

wobei

$$\widehat{\text{erwartet}}_{hk} = n\widehat{\pi}_{h+}\widehat{\pi}_{+k} = N_{h+}N_{+k}/n.$$

χ^2 – *Test*

$$T = \sum_{h,k} \frac{(\text{beobachtet}_{hk} - \widehat{\text{erwartet}}_{hk})^2}{\widehat{\text{erwartet}}_{hk}},$$

Unter H_0 ist T approximativ χ^2-verteilt mit $(r-1)(s-1)$ Freiheitsgraden (unter allen obigen Sampling Methoden).

Achtung: Damit die Approximation vernünftig ist, sollte erwartet_{hk} nicht zu klein sein.

Faustregel: 4/5 der $\widehat{\text{erwartet}}_{hk}$ sollten ≥ 4 sein, die übrigen ≥ 1.

Association Plot

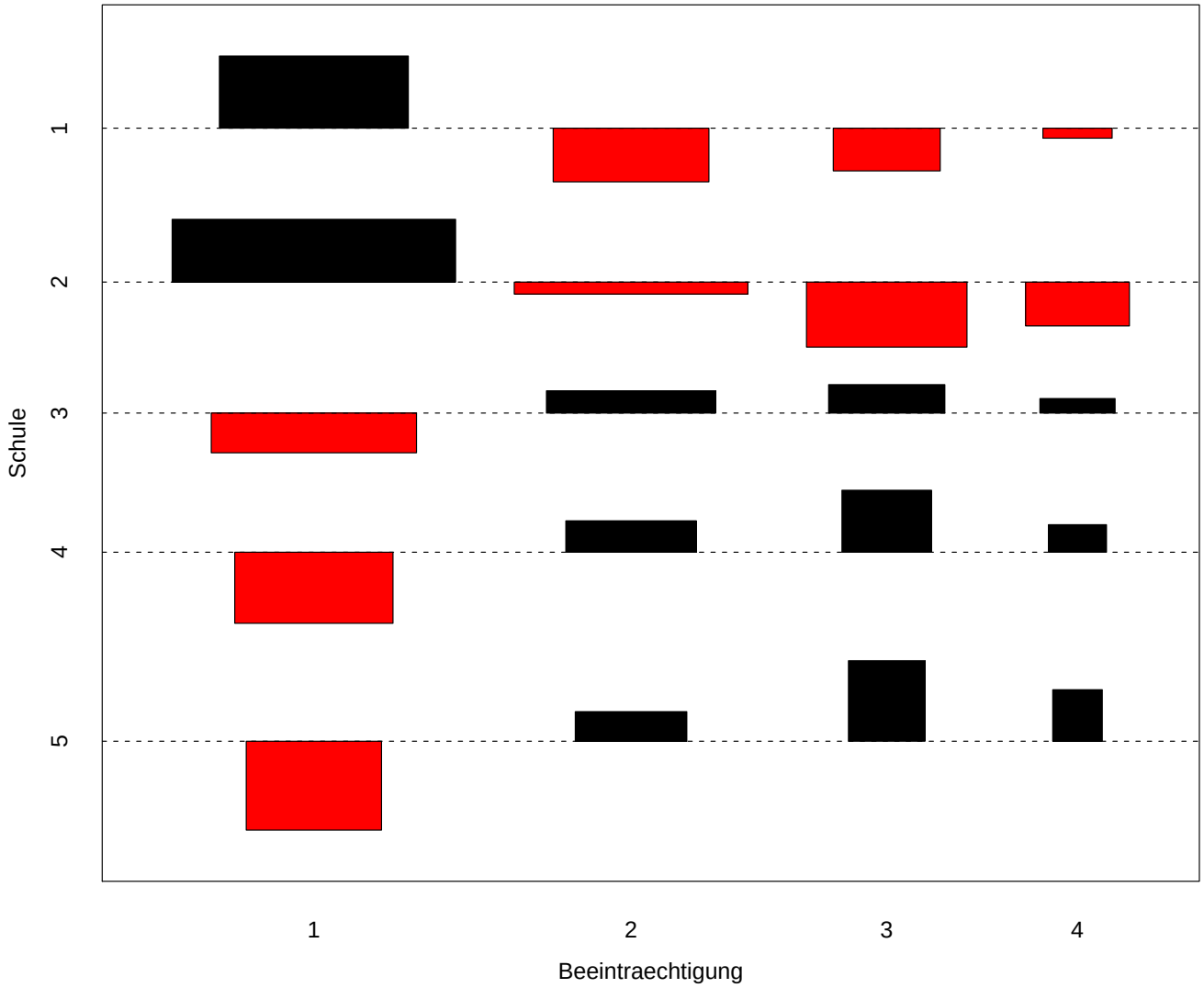


Abbildung: Association Plot für das Beispiel der Umfrage.

Vergleich von unabhängigen Stichproben

Frage: Antworten die Personen mit verschiedener Schulbildung unterschiedlich auf die Frage nach der Belästigung?

Der Test zum Vergleich von unabhängigen Stichproben ist mit dem Test für die Unabhängigkeit zweier Variablen identisch!

Bemerkung

- Quantitative Variable: „Lageparameter“ (Erwartungswert oder Median) von Interesse.
- Für kategorielle Variablen: Vergleich der ganzen Verteilungen!

Vierfeldertafeln ($r = s = 2$)

Beispiel Herzinfarkt und Verhütungsmittel (Agresti, 1990).

58 verheiratete Herzinfarkt-Patientinnen < 45 J. 2 Spitalregionen. Vergleich mit Pat., die aus anderen Gründen ins Spital kamen.

		Herzinfarkt (B)		
		ja	nein	Summe
Verhütungspille	ja	23	34	57
(A)	nein	35	132	167
	Summe	58	166	224

Tabelle: Kreuztabelle der Verwendung von Verhütungspillen und Herzinfarkt.

Ist $N_{11}/n_{+1} = 23/58 = 40\%$ signifikant von
$N_{12}/n_{+2} = 34/166 = 20\%$ verschieden?

Vergleich zweier Wahrscheinlichkeiten (2 Stichproben)
Verwende gleichen Test wie früher. Die Teststatistik kann man umformen zu

$$T = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} .$$

```
Pearson's Chi-squared test with Yates' cont.corr.
X-squared = 7.3488, df = 1, p-value = 0.00671
```

Exakte Verteilung von T hier bestimmbar:
$T \mid n_{1+}, n_{2+}, n_{+1}, n_{+2}$ = Funktion von N_{11}.

$$\begin{aligned} P(N_{11} = n_{11}) &= \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} = \frac{n_{1+}!}{n_{11}!n_{12}!} \cdot \frac{n_{2+}!}{n_{21}!n_{22}!} \Big/ \frac{n!}{n_{+1}!n_{+2}!} \\ &= \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!} \end{aligned}$$

Hypergeometrische Verteilung $\rightsquigarrow$ **Exakter Test von Fisher**.
Fisher-Test liefert P-Wert von 0.00519.

Häufige Fehler

- **Daten falsch aufgeschlüsselt**
  Bsp. 2 von 10 vs. 8 von 12.
  Richtige Tabelle ist $\begin{array}{|cc|}\hline 2 & 8 \\ 8 & 4 \\ \hline\end{array}$ und nicht $\begin{array}{|cc|}\hline 2 & 8 \\ 10 & 12 \\ \hline\end{array}$!
  Fisher Test p-Werte: 0.042 vs. 0.25.
- Keine Anzahlen von unabhängigen Beobachtungen.
  Bsp. Anzahl Pflanzenarten auf Probeflächen.
- Klassen mit zu kleinen Erwartungswerten zusammenfassen, nicht weglassen.

Statistik-Programme

- Daten in Form der üblichen Datenmatrix
- Zeilen entsprechen Beobachtungen $i \rightsquigarrow A_i$, B_i.
- Die Kreuztabelle mit den N_{hk} erstellt das Programm selbst.
- Kreuztabelle direkt eingeben – manchmal unmöglich.
- Nur Anzahlen bekannt $\rightsquigarrow$ eine Zeile pro Kombination $[h, k]$

A	B	N
1	1	23
1	2	35
2	1	34
2	2	132

N: „Gewicht“.

Vierfeldertafel: Verbundene Stichproben

Bsp. aus Rice (1995)
Bei Patienten mit Lymphknotenvergrösserung und einer Kontrollgruppe wurde vermerkt, ob die Mandeln entfernt wurden in der Vergangenheit $\rightsquigarrow$ normale Vierfeldertafel.

In einer Folgestudie wurden 85 Patienten so ausgewählt, dass man von jedem Patienten ein Geschwister hatte, das *nicht* von der Krankheit betroffen war. Wieder hat man die Angabe über die Mandelentfernung.

- 2 (binäre) Variablen $Y^{(1)}$, $Y^{(2)}$
- Ist der Anteil der Personen mit Mandelentfernungen in beiden Populationen gleich?
- 85 Patienten, 85 Geschwister
  $\rightsquigarrow$ Sample size? Was ist hier eine unabhängige Beobachtungseinheit?

Richtige Sichtweise

		Sibling	
		Entfernt	Nicht entfernt
Patient	Entfernt	26	15
	Nicht entfernt	7	37

D.h. wir haben 85 (unabhängige) Beobachtungen.

Falsch wäre die Betrachtung

	Entfernt	Nicht entfernt
Sibling	33	52
Patient	41	44

Denn wir haben *nicht* 170 unabhängige Beobachtungen!

Sind die Verteilungen in beiden Gruppen gleich?

McNemar-Test

$H_0 : \pi_{1+} = \pi_{+1} \iff \pi_{12} = \pi_{21}.$

Unter H_0 gilt: $N_{12} \sim \mathcal{B}(N_{12} + N_{21}, 1/2)$.

Bedingte Verteilung der Anzahl Wechsel von 1 nach 2, gegeben die Anzahl aller Wechsel.

N_{11} und N_{22} „egal“!

```
> binom.test(7, 22, 0.5)

data:  7 and 22
number of successes = 7, number of trials = 22, p-value = 0.1338
alternative hypothesis: true probability of success is not equal to 0.5
```

Vgl. auch Vorzeichentest.

Abhängigkeit von zwei Variablen

Setup

- Binärer Antwortfaktor B ($B = 1$: krank, $B = 2$: gesund), Gruppierungsfaktor A.
- Abhängigkeit durch eine Zahl charakterisieren, die die Stärke des Zusammenhangs misst $\rightsquigarrow$ „Korrelation“.

Definition Risiko

$\pi_{1|h} = P(B = 1 \mid A = h) = \pi_{h1}/\pi_{h+}$ für die Gruppe h.

Vergleich von Risiken

- Risiko-Differenz: $\pi_{1|1} - \pi_{1|2}$.
- **relatives Risiko**: $\pi_{1|1}/\pi_{1|2}$.

Nützlicher ist das Doppelverhältnis.

Doppelverhältnis, Odds Ratio

Wettverhältnis (odds): $\pi_{1|1}/\pi_{2|1}$.

„Chancen“ für $B=1$ in der Gruppe $A=1$.
odds $= 3 \Leftrightarrow \; P(B=1 \mid A=1) = 0.75$.

Doppelverhältnis (odds ratio)

Vergleich der Wettverhältnisse für $A=1$ und $A=2$

$$\theta = \frac{P(B=1 \mid A=1)}{P(B=2 \mid A=1)} \Big/ \frac{P(B=1 \mid A=2)}{P(B=2 \mid A=2)} = \frac{\pi_{1|1}}{\pi_{2|1}} \Big/ \frac{\pi_{1|2}}{\pi_{2|2}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \,.$$

Verhältnis von Verhältnissen $\rightsquigarrow$ **Doppelverhältnis (odds ratio)**.

Zwei Gruppen (Vierfeldertafel) $\rightsquigarrow$ A und B vertauschbar.

Symmetrisches Mass für die Abhängigkeit von zwei binären Variablen.

Eigenschaften

- $\theta = 1 \iff$ bedingte Wahrscheinlichkeiten gleich
  Unabhängigkeit von A und B!
- $\theta > 1$, $r = s = 2$: $\pi_{11} \cdot \pi_{22} >$ als unter Unabhängigkeit.
  „Positive Abhängigkeit“, d.h. W'keit, dass beide Variablen den gleichen Wert annehmen ist erhöht.
  Analog $\theta < 1$.
- θ hängt **nicht** von Randverteilungen ab!
  Geschichtete Stichproben: Doppelverhältnisse richtig!
- Schätzung:

$$\widehat{\theta} = \frac{(N_{11} + 0.5)\,(N_{22} + 0.5)}{(N_{12} + 0.5)\,(N_{21} + 0.5)}.$$

Logarithmiertes Doppelverhältnis (log odds ratio) $\ell\theta = \log(\theta)$

- $\ell\theta = 0$ bei Unabhängigkeit,
- $\ell\theta > 0$ bei positiver Abhängigkeit,
- $\ell\theta < 0$ bei negativer Abhängigkeit.
- Vertauscht man die Kategorien (1 und 2) der einen Variablen, so wechselt nur das Vorzeichen von $\ell\theta$!
- Achtung: $\ell\theta$ nicht auf $[-1, 1]$ begrenzt!

Mehr als zwei Klassen.
Betrachte z.B. nur zwei Gruppen miteinander:

$$\begin{aligned}\theta_{hk,h'k'} &= \frac{P(B=k \mid A=h)}{P(B=k' \mid A=h)} \Big/ \frac{P(B=k \mid A=h')}{P(B=k' \mid A=h')} \\ &= \frac{\pi_{k|h}}{\pi_{k'|h}} \Big/ \frac{\pi_{k|h'}}{\pi_{k'|h'}} = \frac{\pi_{hk}\pi_{h'k'}}{\pi_{h'k}\pi_{hk'}}.\end{aligned}$$

Unabhängigkeit $\iff \theta_{hk,h'k'} = 1$ für alle Kombinationen, d.h. alle Doppelverhältnisse sind 1.

Merkpunkte

- Aus kategoriellen Daten entstehen durch Tabellieren **Häufigkeitsdaten**.
- **Grundlegendes Modell** für Häufigkeitsdaten:
  Unabhängige Poisson-Verteilung.
  $\rightsquigarrow$ bedingte Verteilung, gegeben Randsummen.
  Das Wichtige am Modell: Annahmen über die π_{ij}.
- **Unabhängigkeit** von zwei Merkmalen: **Chiquadrat-Test**.
  Die einzelnen Beiträge (Residuen) können bei der Interpretation eines signifikanten Resultats helfen.
- Abhängigkeitsmass: **Doppelverhältnis**, meist logarithmiert.

Studien

- **Retrospektive Studie (case control study)**
  Bsp. Herzinfarkt. Das absolute Risiko kann nicht geschätzt werden, aber sehr wohl die Doppelverhältnisse!
- **Querschnittstudie (cross sectional study)**
  Ziehe Zufallsstichprobe aus Bevölkerung, kann absolutes Risiko messen (für verbreitete Krankheiten).

Prospektive Studien

- **Kohorten Studie (cohort study)**
  Verfolge Entwicklung von grosser Gruppe (Kohorte). Schaue, bei wem die Krankheit ausbricht, vergleiche mit Ausgangslage.
- **Klinische Studien (clinical trials)**
  Teile Patienten zufällig in Behandlungs- und Kontrollgruppe ein.

Simpson Paradoxon

2-dim Kreuztabelle sagt über Abhängigkeit so viel aus wie einfache Regression: In der Regel **zu wenig!**

Beispiel Zulassung zum Studium

	Anzahlen			Prozente		
Geschl.	zugel.	abgew.	$\sum$	zugel.	abgew.	$\sum$
w	557	1278	1835	30.4	69.6	100
m	1198	1493	2691	44.5	55.5	100
$\sum$	1755	2771	4526	38.8	61.2	100

Diskriminierung!!!

Zulassungen aufgeschlüsselt nach Departement

Dept.	Geschl.	Anzahlen zugel.	Anzahlen abgew.	Anzahlen $\sum$	Prozente zugel.	Prozente abgew.	Prozente $\sum$
A	w	89	19	108	82.4	17.6	100
	m	512	313	825	62.1	37.9	100
B	w	17	8	25	68.0	32.0	100
	m	353	207	560	63.0	37.0	100
C	w	202	391	593	34.1	65.9	100
	m	120	205	325	36.9	63.1	100
D	w	131	244	375	34.9	65.1	100
	m	138	279	417	33.1	66.9	100
E	w	94	299	393	23.9	76.1	100
	m	53	138	191	27.7	72.3	100
F	w	24	317	341	7.0	93.0	100
	m	22	351	373	5.9	94.1	100
	$\sum$	1755	2771	4526	38.8	61.2	100

Ausblick

- Zusammenhänge innerhalb von verschiedenen Gruppen $\neq$ Zusammenhänge ohne Gruppierung!
- Phänomen bekannt unter dem Namen **Simpson's Paradoxon**.
- Regression: Koeffizient eines Regressors kann Vorzeichen wechseln, wenn andere erklärende Variablen ins Modell kommen.
- Bedeutung der Koeffizienten hängt vom Modell ab!
- Mehrere kategorielle Variablen modellieren: **Loglineare Modelle**.