

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

„Disclaimer“: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 bedingten Häufigkeiten sehr viele Fehlerquellen mit sich bringt und große Schwierigkeiten macht.**

→ **Rüstzeug,**

- **damit man selber solche Fehler nicht macht und**
- **Veröffentlichungen kritisch hinterfragen kann.**

Bsp. 5.13.

A. Quatember (Institut für angewandte Statistik, Linz): Unsinn in in den Medien - Vom allzu sorglosen Umgang mit Daten (I):
Man nehme kritisch Stellung zu dem folgenden Zeitungsausschnitt!



Quelle: Kronen-Zeitung, 15.07.2000

Bsp. 5.14.

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

Wiens Schüler fallen öfter durch

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.

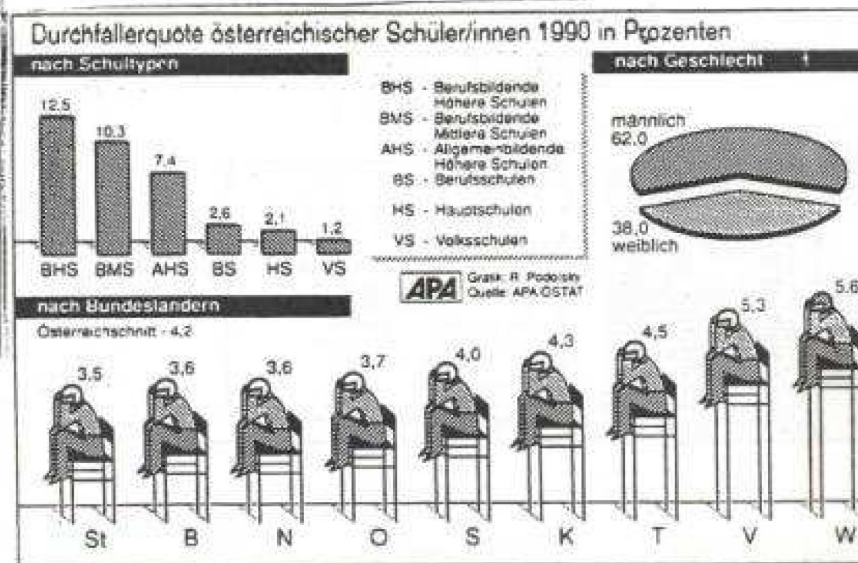
Die Ursachen ortet das ÖSTAT in der unterschiedlichen Leistungsbeurteilung in den Ländern. Im Bundesdurchschnitt dürfen jährlich 4,2 Prozent der Schüler nicht „aufsteigen“.

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)

DER STANDARD:
8.5.1992



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

Bsp. 5.15. *Illustration an dem fiktiven Beispiel (Aggression und Fahrzeugklasse)*

Empirische Unabhängigkeit: Definition

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

$$f_{Y|X}(b_j|a_i) = f_{\bullet j} = f_Y(b_j) \quad (5.14)$$

und

$$f_{X|Y}(a_i|b_j) = f_{i\bullet} = f_X(a_i) \quad (5.15)$$

gilt.

Satz:

- a) Es genügt, entweder (5.14) oder (5.15) zu überprüfen: Mit einer der beiden Beziehungen gilt auch die andere.
- b) X und Y sind genau dann empirisch unabhängig, wenn für alle $i = 1, \dots, k$ und alle $j = 1, \dots, m$ gilt:

$$f_{ij} = f_{i\bullet} \cdot f_{\bullet j}. \quad (5.16)$$

c) Gleichung (5.16) ist äquivalent zu

$$h_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}. \quad (5.17)$$

Beweis:

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

$$\chi^2 := \sum_{i=1}^k \sum_{j=1}^m \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} \quad (5.18)$$

$$= \sum_{\text{alle Zellen}} \frac{(\text{beob. Häufigk.} - \text{unter Unabh. zu erwartende Häufigk.})^2}{\text{unter Unabh. zu erwartende Häufigk.}}$$

Bsp. 5.16.

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

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

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

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 (5.18) 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.

Bem. 5.17.

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

$$\chi^2 = \frac{n \cdot (h_{11}h_{22} - h_{12}h_{21})^2}{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}} \quad (5.19)$$

Veranschaulichung der Formel:

5.3.3 χ^2 -basierte Maßzahlen

Bem. 5.18.

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 für sich genommen ist χ^2 hingegen problematisch und nicht direkt interpretierbar, da sein Wert vom Stichprobenumfang n und von der Zeilen- und Spaltenzahl abhängt \implies 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:

$$K := \sqrt{\frac{\chi^2}{n + \chi^2}}. \quad (5.20)$$

- Zahl liegt immer zwischen 0 und 1 ($0 \leq \text{Nenner} < \text{Zähler}$)

b) Korrigierter Kontingenzkoeffizient:

$$K^* := \frac{K}{K_{\max}} \quad (5.21)$$

mit

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

c) Kontingenzkoeffizient nach Cramér (Cramér's V):

$$\begin{aligned} V &= \sqrt{\frac{\chi^2}{n \cdot (\min\{k, m\} - 1)}} \\ &= \sqrt{\frac{\chi^2}{\text{prinzipiell maximal möglicher Wert von } \chi^2}} \end{aligned} \quad (5.22)$$

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 (5.19) ergibt sich also

$$\Phi = \left| \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}}} \right|. \quad (5.23)$$

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 (Achtung, bei kategorialen Merkmalen auf Kodierung achten.):

$$\Phi_s > 0$$

und

$$\Phi_s < 0$$

Bem. 5.19.

- 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.).

Bsp. 5.20. Berechnung von gängigen χ^2 -basierten Maßzahlen

Beschäftigungsstatus und Geschlecht.

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

• $K =$

• $K_{max} =$

• $K^* =$

• $V =$

• $\Phi_s =$

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

Bsp. 5.21. *Wahlabsicht und Bildungsabschluss (ALLBUS 2010: V327, V747)*



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

Bsp. 5.22. *Beispiel 3: (Aggression und Fahrzeugtyp)*

Fahrzeugtyp & Fahrverhalten

| | | Aggressivität | |
|----------|--------------|---------------|-------------|
| | | aggro | nicht aggro |
| Fahrzeug | Kompakt | | |
| | Mittelklasse | | |
| | Oberklasse | | |

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

Bem. 5.23. *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*

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

Bsp. 5.24. Berechnung von Φ_{korrr})

| X^Y | 1 | 2 | |
|-------|----|----|--|
| 1 | 40 | 25 | |
| 2 | 80 | 5 | |
| | | | |

Vergleich von Maßzahlen über:

- Skalenniveaus der Merkmale
- Grundidee/Konstruktion
- Größe der Kontingenztafeln
- Wertebereich (Interpretation)

χ^2 : χ^2 -Abstand / χ^2 -Koeffizient

Vergleich mit Unabhängigkeitstafel, Summe über (quadrierte, normierte) Abweichungen in jeder Zelle

Wertebereich 0 bis $n \cdot (\min\{k, m\} - 1)$, müsste also immer Obergrenze ausrechnen um einzuschätzen, wie stark der Zusammenhang ist. Obergrenze hängt von n und Anzahl Zeilen/Spalten ab.

0 heißt Unabhängigkeit, je größer desto stärkerer Zusammenhang.

K : Kontingenzkoeffizient nach Pearson

χ^2 normiert auf Anzahl Beobachtungen (n in der Formel), also vergleichbar für

Tabellen mit gleichem „Format“.

Wertebereich 0 bis K_{\max} (prinzipiell bis 1, falls große Anzahl Zeilen/Spalten)

0 heißt Unabhängigkeit, je größer (näher bei K_{\max}) desto stärkerer Zusammenhang.

K^* : *Korrigierter Kontingenzkoeffizient*

K normiert mit K_{\max} , also vergleichbar für beliebige Kontingenztafeln.

Wertebereich 0 bis 1.

0 heißt Unabhängigkeit, je näher bei 1 desto stärkerer Zusammenhang.

V : *Kontingenzkoeffizient nach Cramer / Cramers V*

χ^2 direkt normiert mit χ_{\max}^2 im Nenner, dann noch Wurzel gezogen.

Wertebereich 0 bis 1.

0 heißt Unabhängigkeit, je näher bei 1 desto stärkerer Zusammenhang.

ϕ : *Phi-Koeffizient*

Nur für 4-Felder-Tafeln, entspricht Cramer's V .

Wertebereich 0 bis 1.

0 heißt Unabhängigkeit, je näher bei 1 desto stärkerer Zusammenhang.

ϕ_s : *Signierter Phi-Koeffizient / Punkt-Korrelationskoeffizient*

ϕ ohne Betragsstriche, damit Richtung des Zusammenhangs erkennbar.

Wertebereich -1 bis +1

0 heißt Unabhängigkeit, je näher bei -1 oder +1 desto stärkerer Zusammenhang.

ϕ_{korr} : *Korrigierter Phi-Koeffizient*

ϕ normiert mit ϕ_{extrem} , da Obergrenze 1 nicht unter allen Randverteilungen erreicht werden kann.

Wertebereich 0 bis 1 (1 kann jetzt bei allen 4-Felder-Tafeln wirklich erreicht werden).

0 heißt Unabhängigkeit, je näher bei 1 desto stärkerer Zusammenhang.

ACHTUNG: Bei gegebenen Randverteilungen ist die Obergrenze 1 bei V (und ϕ) eventuell nicht möglich (Grund: eine gedachte Tafel mit perfektem Zusammenhang lässt sich dann nicht mehr konstruieren).

5.4 Weitere Methoden für Vierfeldertafeln

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

| | | Y | |
|-----|------------------|-------|-------------------|
| | | ja | nein |
| | | b_1 | b_2 |
| X | exponiert: | a_1 | h_{11} h_{12} |
| | nicht exponiert: | a_2 | h_{21} h_{22} |

In der Medizin ist das Ereignis meist eine bestimmte Erkrankung. Man bezeichnet dann die bedingte relative Häufigkeit $f_{Y|X}(b_j|a_i)$ als *Risiko* für b_j unter Bedingung a_i :

$$R(b_j|a_i) := f_{Y|X}(b_j|a_i) = \frac{h_{ij}}{h_{i\bullet}} \quad i, j = 1, 2.$$

In der Epidemiologie wird standardmäßig $R(b_1|a_1)$ betrachtet. Dabei sind b_1 und a_1 so gewählt, dass sich das Erkrankungsrisiko für Personen, die exponiert waren, ergibt.

Als Zusammenhangsmaß zwischen X und Y in Vierfelder-Tafeln verwendet man oft das darauf aufbauende *relative Risiko*:

5.4.1 Relatives Risiko und Prozentsatzdifferenz

Definition: Für eine Vierfelder-Tafel heißt

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

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 1 | nein 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

$$d\%(b_j) := (f_{Y|X}(b_j|a_1) - f_{Y|X}(b_j|a_2)) \cdot 100, \quad j = 1, 2$$

heißt *Prozentsatzdifferenz* für b_j .

Bem. 5.25. *Eigenschaften von $d\%$*

- $d\%(b_1)$ ist z.B. die Differenz aus den Ereignisrisiken 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:

Offensichtlich gilt bei zwei Ausprägungen

$$\begin{aligned}d\%_0(b_1) &= (f_{Y|X}(b_1|a_1) - f_{Y|X}(b_1|a_2)) = \\ &= (1 - f_{Y|X}(b_2|a_1)) - (1 - f_{Y|X}(b_2|a_2)) \\ &= -(f_{Y|X}(b_2|a_1)) - f_{Y|X}(b_2|a_1) = \\ &= -d\%_0(b_2)\end{aligned}$$

Bsp. 5.26. *Beschäftigung von Männern und Frauen*

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

Bem. 5.27.

- Den in diesem Abschnitt betrachteten Maßzahlen ist gemeinsam, 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 – wie immer bei zwei Zahlen auch – 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)

⇒ die Differenz ist > 0 , wenn der erste Term größer ist als der zweite.

⇒ die Differenz ist < 0 , wenn der erste Term kleiner ist als der zweite.

- Bei kleinen Risiken ist die Prozentsatzdifferenz nicht sensitiv, z.B.:

$$* f_{Y|X}(b_1|a_1) = 0.42, f_{Y|X}(b_1|a_2) = 0.41$$

$$RR(b_1) = 1.02$$

$$d\%(b_1) = 1$$

$$* f_{Y|X}(b_1|a_1) = 0.02, f_{Y|X}(b_1|a_2) = 0.01$$

$$RR(b_1) = 2.0$$

$$d\%(b_1) = 1$$

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

5.4.2 Odds Ratio

Definition: Die Größe

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

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

Bem. 5.28. *Eigenschaften des Odds Ratio*

- 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:

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

- 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”.

Bsp. 5.29. *Beschäftigung von Männern und Frauen*

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

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

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

Genau wie ein einzelnes 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.

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).

Bem. 5.30. *Eigenschaften des Odds Ratio*

- 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 reduzieren, wird gelegentlich OR logarithmiert, also $\ln OR$ betrachtet. Sein Wertebereich ist $(-\infty, \infty)$, wobei nun der

Wert 0 auf keinen Zusammenhang hinweist. Dennoch ist der Abstand zu 0 zweier analoger Situationen nicht gleich.

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

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

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.

Bsp. 5.31. *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* .

Bem. 5.32.

- 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 $Q = -1$, und Q zeigt also dann bereits eine perfekte Abhängigkeit.

Bsp. 5.33. *Beschäftigung von Männern und Frauen*

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