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 Schichtzugehörigkeit)

X \ Y	1	2	
1	2	2	4
2	1	1	2
3	5	1	6
	8	4	12

X \ Y	1	2	
1			
2			
3			

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:

Beurteilung des Ausmaßes der Abhängigkeit von X und Y. Durch Vergleich der beobachteten Kontingenztafel mit der Tafel, die sich bei denselben Randverteilungen ergeben würde, wenn X und Y (empirisch) unabhängig wären.

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

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

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 $\tilde{h}_{ij}$.

$$\tilde{h}_{11} = \frac{h_{1\bullet} \cdot h_{\bullet 1}}{n} = \frac{65 \cdot 120}{150} = 52 \qquad \tilde{h}_{12} =$$

$$\tilde{h}_{21} = \qquad \tilde{h}_{22} =$$

bei empirischer Unabhängigkeit zu erwartende Kontingenztafel:

X \ Y	1	2	
1	52		65
2			85
	120	30	150

Man erhält

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(h_{ij}-\tilde{h}_{ij})^2}{\tilde{h}_{ij}} \\
&= \frac{(40-52)^2}{52} + \\
&= \\
&=
\end{aligned}$$

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:

In der Hauptdiagonalen stehen die „gleichgerichteten (konkordanten) Paare“ (1,1), (2,2), in der Nebendiagonalen die „entgegengerichteten (diskordanten) Paare“ (1,2), (2,1). Je stärker der Zusammenhang, desto größer ist $(h_{11}h_{22} - h_{12}h_{21})^2$.

Berechnung im Beispiel mit alternativer Formel:

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

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 vorkommen. Dies entspräche dann einem perfektem Zusammenhang.

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):

$$\begin{aligned} V &= \sqrt{\frac{\chi^2}{n \cdot (\min\{k,m\} - 1)}} \\ &= \sqrt{\frac{\chi^2}{\text{maximaler Wert}}} \end{aligned} \tag{5.9}$$

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

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

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

Korrekturverfahren für Φ (normiere auf den maximal möglichen Wert bei den gegebenen Randverteilungen; vgl. Wagschal (1999))

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

   i. Berechne das Vorzeichen von Φ_s:
   Ist $h_{11}h_{22} - h_{12}h_{21} > 0$, so setze $min(h_{12}, h_{21})$ auf 0.
   Ist $h_{11}h_{22} - h_{12}h_{21} < 0$, 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 Phi-Koeffizienten* bzw. den zugehörigen *korrigierten signierten Phi-Koeffizienten*

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

Berechnung im Beispiel:

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

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

Mit (5.10) erhält man

$$\Phi_{extrem} =$$

und damit

$$\Phi_{korr} =$$

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.

$$R(\text{beschäftigt}|\text{Frau}) = f(b_1|a_1) = \frac{h_{11}}{h_{1\bullet}} =$$

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

$$=$$

$$RR(\text{arbeitslos}) = \frac{R(b_2|a_1)}{R(b_2|a_2)} = \frac{h_{12}/h_{1\bullet}}{h_{22}/h_{2\bullet}}$$

$$=$$

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.

$d\%(b_1)$ ist z.B. die Differenz aus den Erkrankungsrisiken für Personen, die exponiert waren, und für Personen, die nicht exponiert waren.

Eigenschaften: $d\%(b_j)$ 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

$$
\begin{aligned}
d\%(b_1) &= (f(b_1|a_1) - f(b_1|a_2)) \cdot 100 \\
&= \left(\frac{h_{11}}{h_{1\bullet}} - \frac{h_{21}}{h_{2\bullet}}\right) \cdot 100 \\
&=
\end{aligned}
$$

$$
d\%(b_2) =
$$

Offensichtlich gilt bei zwei Ausprägungen

$$\begin{aligned} d\%(b_1) &= (f(b_1|a_1) - f(b_1|a_2)) \cdot 100 = \\ &= (1 - f(b_2|a_1)) \cdot 100 - (1 - f(b_2|a_2)) \cdot 100 \\ &= -(f(b_2|a_1) - f(b_2|a_2)) \cdot 100 = \\ &= -d\%(b_2) \end{aligned}$$

Bemerkungen:

- Bei den in diesem Abschnitt betrachteten Maßzahlen verändert das Vertauschen von Zeilen und Spalten die Maßzahl (im Gegensatz zu den χ^2-basierten Maßzahlen). 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. Vergleichswert ist 1)

    $\rightarrow$ 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. Vergleichswert ist 0)

    $\rightarrow$ 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 zu 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 die Log-Odds-Ratio $\ln OR$ betrachtet. Ihr Wertebereich ist $(-\infty, \infty)$, wobei nun der Wert 0 auf keinen Zusammenhang hinweist.

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

$$OR(b_1) := \frac{O(b_1|a_1)}{O(b_1|a_2)} = \frac{\frac{R(b_1|a_1)}{1-R(b_1|a_1)}}{\frac{R(b_1|a_2)}{1-R(b_1|a_2)}} = \frac{\frac{f(b_1|a_1)}{f(b_2|a_1)}}{\frac{f(b_1|a_2)}{f(b_2|a_2)}}$$

$$= \frac{\frac{h_{11}/h_{1\bullet}}{h_{12}/h_{1\bullet}}}{\frac{h_{21}/h_{2\bullet}}{h_{22}/h_{2\bullet}}} = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11} \cdot h_{22}}{h_{21} \cdot h_{12}}$$

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

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

Die Differenz im Zähler

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

wird groß, 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

$$h_{11}h_{22} \gg h_{12}h_{21} \quad \text{oder} \quad h_{11}h_{22} \ll h_{12}h_{21}$$

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

In der 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 die 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

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

5.4.3 Yules Q

Definition: Die Größe

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

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

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

$$=$$

5.5 PRE-Maße (Prädiktionsmaße)

5.5.1 Die grundlegende Konstruktion

- Völlig andere, sehr allgemeine Grundidee zur Beschreibung von Zusammenhängen.
- Grundlegendes Prinzip vieler statistischer Konzepte.
- Hängt mit Streuungszerlegung metrischer Daten zusammen.
- Anwendbar für *Kreuztabellen beliebiger Größe*.
- In der Soziologie sehr gebräuchlich, da das Prinzip auf sehr viele unterschiedliche Situationen anwendbar ist.

Hintergrund: Naiv ausgedrückt, versucht ein „Modell“ ein empirisches Phänomen zu beschreiben. Ein Modell ist dann umso „besser“, je genauer es ein Phänomen reproduzieren/vorhersagen kann. Die Vorhersagekraft der einen Variablen für die andere dient dann als Maß des Zusammenhangs.

Betrachte zwei Modelle (Modell 1 und Modell 2) zur Vorhersage des Wertes y_i der abhängigen Variable Y einer beliebigen Beobachtung i, wobei Modell 2 die Informationen von Modell 1 und weitere Informationen benutzt.

Anwendung bei der Analyse von Kreuztabellen :

Modell 1: verwendet (ausschließlich) die Randverteilung von Y: $(h_{\bullet j}), j = 1, \ldots, m$.

Modell 2: verwendet die $gemeinsame$ Verteilung von (X, Y) bzw. die bedingte Verteilung von Y gegeben X.

Definition: PRE = Proportional Reduction in Error

$$PRE = \frac{E_1 - E_2}{E_1} = 1 - \frac{E_2}{E_1}$$

wobei

$$E_1: \quad \text{Vorhersagefehler bei Modell 1}$$

$$E_2: \quad \text{Vorhersagefehler bei Modell 2}$$

PRE ist auf $[0, 1]$ normiert, da die Modelle so konstruiert sind, dass automatisch $E_2 \leq E_1$ gilt:

- $PRE = 1$ gilt genau dann wenn $E_2 = 0$, d.h. bei vollständiger Vorhersage bzw. vollständigem Zusammenhang.
- $PRE = 0$ gilt genau dann wenn $E_1 = E_2$, d.h. die Vorhersage wird durch Kenntnis der unabhängigen Variablen in keinster Weise unterstützt, d.h. es besteht kein Zusammenhang.

Intuitives Beispiel:

Y: Beschäftigungsstatus

X_1: Geschlecht

X_2:

5.5.2 Guttmans Lambda

Basiert auf dem Modus der Randverteilung bzw. der bedingten Verteilungen.
Beispiel: Geschlecht $\rightarrow$ Erwerbsstatus

- Modell 1 (nur Y): Modus der Randverteilung:

  Fehler im Modell 1:

  - Richtig vorhergesagt werden alle Einheiten, deren Ausprägungen tatsächlich auf den Modus fallen, das sind $\max_j(h_{\bullet j})$ Einheiten.
  - Es gilt also $E_1 = n - \max\limits_j(h_{\bullet j})$.

- Modell 2 (mit X): Modus unter der Bedingung $X = a_i$, $i = 1, \ldots, k$.

  Fehler im Modell 2, „bedingte Modi“:

  - Aufspalten nach den einzelnen Werten a_i, $i = 1, \ldots, k$.

– Korrekt vorhergesagt werden jeweils diejenigen Einheiten, deren Ausprägungen tatsächlich auf den „bedingten Modus“ fallen, das sind, für jedes feste i, $\max_j(h_{ij})$ Einheiten.

– Es gilt also

$$E_2 = \sum_{i=1}^{k}(h_{i\bullet} - \max_j(h_{ij})) = n - \sum_{i=1}^{k}\max_j(h_{ij})$$

PRE-Maß für abhängige Variable Y:

$$\begin{aligned}\lambda_Y &= \frac{E_1 - E_2}{E_1} = \frac{\left(n - \max_j(h_{\bullet j})\right) - \left(n - \sum_{i=1}^{k}\max_j(h_{ij})\right)}{n - \max_j(h_{\bullet j})} \\ &= \frac{\left(\sum_{i=1}^{k}\max_j(h_{ij})\right) - \max_j(h_{\bullet j})}{n - \max_j(h_{\bullet j})}\end{aligned}$$

Wenn unklar ist, welche Variable die abhängige und welche die unabhängige ist, dann bildet man eine symmetrische Version. Dazu betrachtet man zunächst analog die Prognose von X (ohne und mit Y). Die entsprechende Formel ergibt sich durch Vertauschen der Rolle von X und Y:

$$\lambda_X = \frac{\left(\sum_{j=1}^{m} \max_i(h_{ij})\right) - \max_i(h_{i\bullet})}{n - \max_i(h_{i\bullet})}$$

Symmetrische Version durch „poolen“:

$$\lambda = \frac{\sum_{i=1}^{k} \max_j(h_{ij}) + \sum_{j=1}^{m} \max_i(h_{ij}) - \max_j(h_{\bullet j}) - \max_i(h_{i\bullet})}{2n - \max_j(h_{\bullet j}) - \max_i(h_{i\bullet})}.$$

Beispiel: Erwerbstätigkeit von Männern und Frauen

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

5.5.3 Goodmans und Kruskals Tau

Idee: statt deterministischer Vorhersagen (immer Modus) probabilistische Vorhersagen (mit Wahrscheinlichkeiten).

1. **Modell 1**: Vorhersage „b_j“ mit Wahrscheinlichkeit $f_{\bullet j}$, $j = 1, \ldots, m$. (z.B. bei einem Beschäftigtenanteil von $2/3$ der Personen nicht immer „Beschäftigung“, sondern im Durchschnitt bei 3 Personen 2-mal „Beschäftigung“ und 1 mal „Arbeitslosigkeit“. Prognose: Auswürfeln mit Wahrscheinlichkeitsverteilung $f_{i\bullet}$, also hier z.B. bei einem Verhältnis $(2/3, 1/3)$:

   - wenn die Augenzahl 1 bis 4 dann Prognose = „Beschäftigung“
   - wenn 5 oder 6 dann Prognose = „Arbeitslosigkeit“

2. **Modell 2**: Für jedes i Vorhersage „b_j“ mit Wahrscheinlichkeit $f(b_j|a_i) = \frac{h_{ij}}{h_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}$, d.h. es werden die relativen Häufigkeiten in den aus X gebildeten Subgruppen eingesetzt.

Man kann zeigen (mit Hilfe der Wahrscheinlichkeitsrechnung, nächstes Semester):

$$\text{erwarteter Wert von } E_1 \quad = \quad 1 - \sum_{j=1}^{m} f_{\bullet j}^2$$

$$\text{erwarteter Wert von } E_2 \quad = \quad 1 - \sum_{j=1}^{m} \sum_{i=1}^{k} \frac{f_{ij}^2}{f_{i\bullet}}$$

Damit ergibt sich:

$$\tau_Y = \frac{\sum_{j=1}^{m} \sum_{i=1}^{k} \frac{f_{ij}^2}{f_{i\bullet}} - \sum_{j=1}^{m} f_{\bullet j}^2}{1 - \sum_{j=1}^{m} f_{\bullet j}^2} \qquad \tau_X = \frac{\sum_{i=1}^{k} \sum_{j=1}^{m} \frac{f_{ij}^2}{f_{\bullet j}} - \sum_{i=1}^{k} f_{i\bullet}^2}{1 - \sum_{i=1}^{k} f_{i\bullet}^2}$$

und die symmetrische Form

$$\tau = \frac{\sum_{j=1}^{m}\sum_{i=1}^{k}\frac{f_{ij}^2}{f_{i\bullet}} + \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{f_{ij}^2}{f_{\bullet j}} - \sum_{j=1}^{m} f_{\bullet j}^2 - \sum_{i=1}^{k} f_{i\bullet}^2}{2 - \sum_{j=1}^{m} f_{\bullet j}^2 - \sum_{i=1}^{k} f_{i\bullet}^2}$$

Definition: Die entsprechenden Größen heißen Goodmans und Kruskals τ_Y, τ_X und τ.

Beispiel: Erwerbstätigkeit von Männern und Frauen

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

In relative Häufigkeiten umrechnen:

	1	2	
1	$\frac{4}{15}$	$\frac{1}{6}$	$\frac{13}{30}$
2	$\frac{8}{15}$	$\frac{1}{30}$	$\frac{17}{30}$
	$\frac{4}{5}$	$\frac{1}{5}$	1

$$
\begin{aligned}
\tau_Y &= \frac{\sum_{j=1}^{m}\sum_{i=1}^{k}\frac{f_{ij}^2}{f_{i\bullet}} - \sum_{j=1}^{m} f_{\bullet j}^2}{1 - \sum_{j=1}^{m} f_{\bullet j}^2} \\
&= \frac{\frac{f_{11}^2}{f_{1\bullet}} + \frac{f_{21}^2}{f_{2\bullet}} + \frac{f_{12}^2}{f_{1\bullet}} + \frac{f_{22}^2}{f_{2\bullet}} - (f_{\bullet 1}^2 + f_{\bullet 2}^2)}{1 - (f_{\bullet 1}^2 + f_{\bullet 2}^2)} \\
&= \frac{\frac{(4/15)^2}{13/30} + \frac{(8/15)^2}{17/30} + \frac{(1/6)^2}{13/30} + \frac{(1/30)^2}{17/30} - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right)}{1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right)} \\
&= \frac{0.732 - \frac{17}{25}}{\frac{8}{25}} \approx 0.1625
\end{aligned}
$$

5.6 Zusammenhangsanalyse bivariater ordinaler Merkmale

Jetzt betrachten wir bivariate Merkmale (X, Y), wobei sowohl X als auch Y (mindestens) ordinales Messniveau aufweisen. Die Ausprägungen von X und Y sind also (in inhaltlich sinnvoller Weise) geordnet.

Beachte: Beide Merkmale müssen ordinal sein, bei einem ordinalen und einem nominalem Merkmal sind Methoden für nominale Merkmale zu verwenden. („Das schwächste Glied in der Kette gibt den Ausschlag!“)

5.6.1 Konkordante Paare

Beispiel: Daten des Schweizer Arbeitsmarktsurvey (aus Jann, 2002, S. 82)

Merkmale:

X: Bildung
Y: Einkommen

jeweils mit den Ausprägungen:

1 niedrig
2 mittel
3 hoch

X \ Y	1	2	3	
1	262	125	8	395
2	496	837	149	1482
3	160	361	268	789
	918	1323	425	2666

Ferner betrachte man die folgenden Einheiten (fiktiv):

Person	Ausprägung von Y Einkommen	Ausprägung von X Bildung
1	3 (hoch)	3 (hoch)
2	2 (mittel)	1 (niedrig)
3	3 (hoch)	2 (mittel)
4	1 (niedrig)	1 (niedrig)
5	2 (mittel)	1 (niedrig)
6	1 (niedrig)	3 (hoch)

Bei Fragen nach Zusammenhängen spielt die $Richtung$ eine Rolle. Man spricht von einem

- $gleichsinnigen\ (gleichläufigen)\ Zusammenhang$, wenn hohe Y-Werte zu großen X-Werten und kleine Y-Werte zu kleinen X-Werten gehören.
- $gegensinnigen\ (gegenläufigen)\ Zusammenhang$, wenn hohe Y-Werte zu niedrigen X-Werten und umgekehrt gehören.

Idee: Zur Messung des Zusammenhangs betrachtet man alle Paare von Einheiten und zählt, wie oft sich ein gleichsinniger und wie oft sich ein gegensinniger Zusammenhang zeigt.

Der Zusammenhang ist umso stärker, je deutlicher eine der beiden „Zusammenhangstendenzen“ überwiegt.

Definition: Gegeben sei die Urliste eines bivariaten Merkmals (X, Y), wobei X und Y jeweils ordinales Skalenniveau besitzen. Ein Paar $(i, j), i \neq j$, von Einheiten mit den Ausprägungen (x_i, y_i) und (x_j, y_j) heißt

a) *konkordant* (gleichläufig), falls entweder

$$(x_i > x_j \text{ und } y_i > y_j)$$

oder

$$(x_i < x_j \text{ und } y_i < y_j)$$

gilt.

Beispiele:

b) $diskordant$ (gegenläufig), falls entweder

$$(x_i > x_j \text{ und } y_i < y_j)$$

oder

$$(x_i < x_j \text{ und } y_i > y_j)$$

gilt.

Beispiele:

c) $ausschließlich\ in\ X\ gebunden$, falls

$$x_i = x_j \text{ und } y_i \neq y_j$$

Beispiel:

d) $ausschließlich\ in\ Y\ gebunden$, falls

$$x_i \neq x_j \text{ und } y_i = y_j$$

Beispiel:

e) $in\ X\ und\ Y\ gebunden$, falls

$$x_i = x_j \text{ und } y_i = y_j$$

Beispiel:

Ferner bezeichne

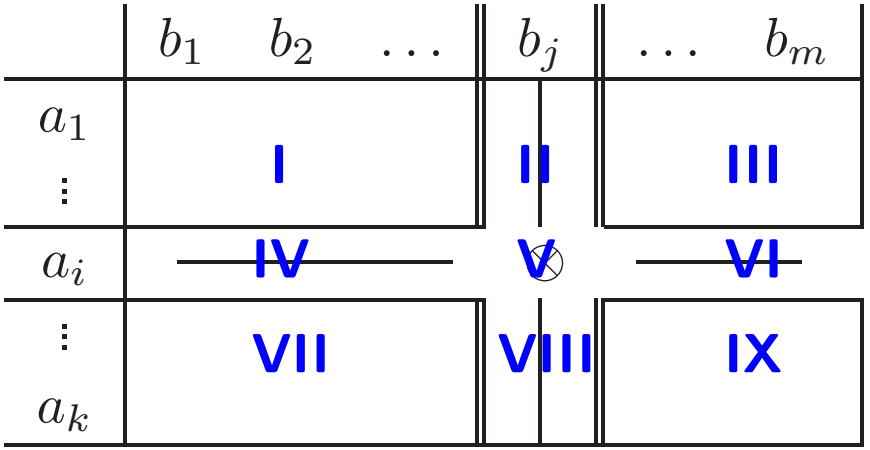
- C die Anzahl der konkordanten Paare,
- D die Anzahl der diskordanten Paare,
- T_X die Anzahl der Paare mit Bindungen ausschließlich in X,
- T_Y die Anzahl der Paare mit Bindungen ausschließlich in Y,
- T_{XY} die Anzahl der Paare mit Bindungen in X und Y.

Die Bezeichnung „T“ kommt vom englischen „Ties“.

Vorsicht: In der Literatur wird manchmal T_{XY} bei T_X und T_Y dazugezählt $\rightarrow$ andere Formeln!

Zur Berechnung geht man die Kreuztabelle Zelle für Zelle durch und zählt jeweils die entsprechenden Paare ab. In jedem Ausprägungspaar (a_i, b_j) lässt sich die Kreuztabelle „zerlegen“.

Sei $a_1 < a_2 < \ldots < a_i < \ldots < a_k$ und $b_1 < b_2 < \ldots < b_j < \ldots < b_m$, dann gilt:



I, IX: Einheiten, die ein konkordantes Paar erzeugen

III, VII: Einheiten, die ein diskordantes Paar erzeugen

V: Einheiten, die ein Paar mit Bindung in X und Y erzeugen = Zellenhäufigkeit - 1

IV, VI: Einheiten, die ein Paar mit Bindung nur in X erzeugen

II, VIII: Einheiten, die ein Paar mit Bindung nur in Y erzeugen

Summiert man die Häufigkeiten auf, so hat man jedes Paar doppelt gezählt, so dass man durch 2 teilen muss. Es gibt schnellere, aber dafür unübersichtlichere Arten zu zählen. (Vorsicht: In der Literatur sind verschiedene Arten zu zählen gebräuchlich.)

Beispiel: Schichtzugehörigkeit und Aggressivität (fiktiv), wobei hier ja/nein als ordinal aufgegefasst wird.

		ja	nein	
		1	2	
Unterschicht	1	2	2	4
Mittelschicht	2	1	1	2
Oberschicht	3	5	1	6
		8	4	12

Zelle (a_i, b_j)	h_{ij}	für C	für D	für T_Y	für T_X	$T_{XY} = h_{ij} - 1$
(1,1)	2		0		2	1
(1,2)	2	0			2	1
(2,1)	1	1	2	7	1	0
(2,2)	1	2	5		1	0
(3,1)	5	0	3	3	1	4
(3,2)	1		0		5	0

Anmerkung zu $T_{XY} = h_{ij} - 1$*:* Zu jeder der h_{ij} Beobachtungen mit Ausprägung (a_i, b_j) gibt es h_{ij} gleiche.

$$
\begin{aligned}
C &= (2 \cdot 2 + 2 \cdot 0 + 1 \cdot 1 + 1 \cdot 2 + 5 \cdot 0 + 1 \cdot 3)/2 = 5 \\
D &= (2 \cdot 0 + 2 \cdot 6 + 1 \cdot 2 + 1 \cdot 5 + 5 \cdot 3 + 1 \cdot 0)/2 = 17 \\
T_Y &= (2 \cdot 6 + 2 \cdot 2 + 1 \cdot 7 + 1 \cdot 3 + 5 \cdot 3 + 1 \cdot 3)/2 = 22 \\
T_X &= (2 \cdot 2 + 2 \cdot 2 + 1 \cdot 1 + 1 \cdot 1 + 5 \cdot 1 + 1 \cdot 5)/2 = 10 \\
T_{XY} &= (2 \cdot 1 + 2 \cdot 1 + 1 \cdot 0 + 1 \cdot 0 + 5 \cdot 4 + 1 \cdot 0)/2 = 12
\end{aligned}
$$

Zur Kontrolle: Insgesamt muss es $\frac{n(n-1)}{2}$ verschiedene Paare geben.

Zusammenhangsmaße für ordinale Daten betrachten nun die (geeignet normierte) Differenz von konkordanten und diskordanten Paaren; sie unterscheiden sich lediglich in der Behandlung von Bindungen und damit in der Normierung.

5.6.2 Zusammenhangsmaße τ_a, τ_b und γ für ordinale Daten

Definition: Die Zusammenhangsmaße für ordinale Daten heißen

$$\tau_a := \frac{C - D}{\frac{n \cdot (n-1)}{2}}$$

Kendalls $Tau\ a$,

$$\tau_b := \frac{C - D}{\sqrt{(C + D + T_X) \cdot (C + D + T_Y)}}$$

Kendalls $Tau\ b$ und

$$\gamma := \frac{C - D}{C + D}$$

Goodmans und Kruskals $Gamma$.

Eigenschaften

- Die Maßzahlen liegen jeweils zwischen -1 und 1.
- Der Zusammenhang ist umso stärker, je größer der Betrag ist. (0: kein Zusammenhang, -1,+1: maximaler Zusammenhang).

- Das Vorzeichen gibt Auskunft über die Richtung des Zusammenhangs:

- Allgemein gilt:

$$|\tau_a| \leq |\tau_b| \leq |\gamma|.$$

  Liegen keine Bindungen vor, sind alle Maßzahlen gleich.

- Bei Bindungen kann τ_a die Extremwerte -1 und 1 nicht erreichen, selbiges gilt bei asymmetrischen Tabellen $(k \neq m)$ für τ_b.

- Die Maßzahlen basieren auf einem etwas unterschiedlichen Verständnis des Begriffs „Zusammenhang“. γ vernachlässigt Bindungen völlig und ist daher ein Maß für die Stärke eines $schwach$ monotonen Zusammenhangs, während τ_a und τ_b sich eher auf $stark$ monotone Zusammenhänge beziehen.

- Wegen der Vernachlässigung von Bindungen reagiert γ sehr sensibel auf das Zusammenfassen von Kategorien.

- γ ist eine Verallgemeinerung von Yules Q. (vgl. Kapitel 5.4.3)

Beispiel: Aggressivität und Schichtzugehörigkeit

Mit den Ergebnissen $C = 5$, $D = 17$, $T_Y = 22$, $T_X = 10$, $n = 12$) ergibt sich

$$\tau_a =$$

$$\tau_b =$$

$$=$$

$$\gamma =$$

Beispiel: Daten des Schweizer Arbeitsmarktsurvey

$$\tau_b = 0.332, \qquad \gamma = 0.533$$

Ähnliche Interpretation, jetzt aber positives Vorzeichen! Einkommen steigt tendenziell mit der Bildung, Bildung wirkt sich jedenfalls im Durchschnitt nicht nachteilig aus.