5.5 PRE-Maße (Fehlerreduktionsmaße)

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 Verbesserung der Modellanpassung der einen Variable durch Berücksichtigung einer (zusätzlichen) anderen Variablen dient dann als Maß des Zusammenhangs zwischen den beiden.

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. Dieses Prinzip wird bei der Analyse von Kreuztabellen wie folgt umgedreht:

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 : Fehler der aus dem Modell 1 abgeleiteten Werte

E_2 : Fehler der aus dem Modell 2 abgeleiteten Werte

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

Konstruktion von PRE-Maßen benötigt also:

- Geeignete Konstruktion eines Fehlermaßes
- Zwei geeignet verschachtelte zu vergleichende Modelle

5.5.2 Guttmans Lambda

Basiert auf dem Modus der Randverteilung bzw. der bedingten Verteilungen.

- Modell 1 (nur Y):

- Modell 2 (mit X):

- Fehler im Modell 1 also:

- Fehler im Modell 2, „bedingte Modi“:

PRE-Maß für abhängige Variable Y:

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

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:

$$\lambda_X = \frac{E_1^X - E_2^X}{E_1^X} = \frac{\left(\sum_{j=1}^{m} \max_i(h_{ij})\right) - \max_i(h_{i\bullet})}{n - \max_i(h_{i\bullet})}$$

Symmetrische Version durch „poolen“:

$$\lambda = \frac{(E_1^X - E_2^X) + (E_1^Y - E_2^Y)}{E_1^X + E_1^Y} = \frac{\sum_{i=1}^{k} \max_j(h_{ij}) + \sum_{j=1}^{m} \max_i(h_{ij}) - \max_j(h_{\bullet j}) - \max_i(h_{i\bullet})}{2n - \max_j(h_{\bullet j}) - \max_i(h_{i\bullet})}.$$

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).

Modell 1: Vorhersage „b_j“ mit Wahrscheinlichkeit $f_{\bullet j}$, $j = 1, \ldots, m$. (z.B. bei einem Beschäftigtenanteil von $2/3$ 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 Würfel 1 bis 4 dann Prognose = „Beschäftigung“
wenn 5 oder 6 dann Prognose = „Arbeitslosigkeit“)

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 (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:

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

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 Beispiel-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)

Da ordinalskalierte Merkmale betrachtet werden, spielt bei Fragen nach Zusammenhängen die *Richtung* eine Rolle. In Verallgemeinerung zu den Überlegungen bei den dichotomen Merkmalen spricht man 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 deutlich 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

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

d) *ausschließlich in Y gebunden*, falls

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

e) *in X und Y gebunden*, falls

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

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 $\Longrightarrow$ scheinbar andere Formeln!

Zur Berechnung geht man die Kreuztabelle Zelle für Zelle durch und zählt jeweils die entsprechenden Paare ab. In jedem Paar von Einheiten mit den Ausprägungen (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:

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

Beispiel: Fahrzeugklasse und Aggression (fiktiv), wobei hier nein/ja als ordinal aufgefasst wird.

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

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

		nein	ja	
		1	2	
Kompaktklasse	1	2	2	4
Mittelklasse	2	1	1	2
Oberklasse	3	1	5	6
		4	8	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	5	2	3	1	0
(2,2)	1	2	1		1	0
(3,1)	1	0	3	3	5	0
(3,2)	5		0		1	4

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

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

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

$$\text{Kendalls } Tau\ a: \qquad \tau_a := \frac{C - D}{\frac{n \cdot (n-1)}{2}}$$

$$\text{Kendalls } Tau\ b: \qquad \tau_b := \frac{C - D}{\sqrt{(C + D + T_X) \cdot (C + D + T_Y)}}$$

$$\text{Goodmans / Kruskals } Gamma: \qquad \gamma := \frac{C - D}{C + D}$$

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: Fahrzeugklasse und Aggression

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

$$\begin{aligned} \tau_a &= \\ \tau_b &= \\ \\ \gamma &= \end{aligned}$$

Beispiel: Daten des Schweizer Arbeitsmarktsurvey

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

Ähnliche Interpretation: Einkommen steigt tendenziell mit der Bildung, Bildung wirkt sich jedenfalls im Durchschnitt nicht nachteilig aus.