Seminar: Dimensionale Analyse und Typenbildung
Seminarleiter: Tilo Görl
Referent: Michael Schüler
Datum: 7.1.2002

Die Korrespondenzanalyse

Die Korrespondenzanalyse ist ein exploratives Verfahren zur grafischen und numerischen Darstellung von Zeilen und Spalten beliebiger Kontingenztabellen. Bei der Korrespondenzanalyse ist die grafische Interpretation zentral. Die manifesten Variablen der Korrespondenzanalyse sind kategorial skaliert.
> "Hauptkomponentenanalyse mit kategorialen Daten"

Analyse-Spektrum:
- Analyse einer Kontingenztabelle mit 2 Variablen
- zusammengesetzte Kontingenztabelle. Eine Spaltenvariable wird mit mehreren Zeilenvariablen kreuztabelliert.
- multiple Korrespondenzanalyse: es wird der Effekt von jeder Variablen auf jede andere Variable berücksichtigt : alle Interaktionen erster Ordnung

Das Verfahren impliziert keine Richtung der Abhängigkeit; diese ist immer Bestandteil der Interpretation. Bei der Korrespondenzanalyse gibt es eine „zu beschreibende Variable" und eine oder mehrere „beschreibende Variablen" – dies ist nicht gleich „abhängig" und „unabhängig", welche die kausalen Abhängigkeiten formulieren.

Bei der zentralen grafischen Darstellung der Ergebnisse der Korrespondenzanalyse gilt es 3 Grundkonzepte (Regeln der Interpretation) zu beachten.
1. Interpretation der Profile
2. Berücksichtigung der Massen
3. Die Chi-Quadrat-Distanzen (gewichtete Euklidische Distanzen)

Ausgangsbeispiel

In einer 1986 durchgeführten Allgemeinen Bevölkerungsumfrage (ALLBUS 1986) wurden Fragen zur kulturellen Kompetenz gestellt. Zum Beispiel: Fähigkeit Hosen umzunähen, Schach zu spielen, nach Popmusik tanzen, Walzer tanzen, etc. .
Antwortmöglichkeiten: ja, kann ich / ja, kann ich etwas / nein, kann ich nicht. (Zeilenvariable)
Als Spaltenvariable wurde eine zusammengesetzte Variable aus Alter und Geschlecht gewählt.

Ausgangsmatrix; absolute Werte
6*3 Kontingenztabelle

	GeschAlt						
HOSE umnähen	M 0-39	m 40-59	m 60 -	w 0-39	w 40-59	w 60 -	Aktiver Rand
Ja	167	109	91	637	500	406	1910
Etwas	132	76	62	32	12	12	326
Nein	308	291	165	23	10	11	808
Aktiver Rand	607	476	318	692	522	429	3044

Interpretation der Profile (1.) und der Begriff der „Masse" (2.)

Es gibt **Spaltenprofile und Zeilenprofile**.

Um in einer Tabelle die Spaltenprofile ablesen zu können, werden die Rohdaten der Kontingenztabelle ins Verhältnis zu den Randsummen (der Spalten) der Variable gesetzt.
Hiermit werden die Spaltensummen auf den Wert 1 normiert.

Spaltenprofile

	GeschAlt						
HOSE	m 0-39	m 40-59	m 60 -	w 0-39	w 40-59	w 60 -	Masse
Ja	,275	,229	,286	,921	,958	,946	,627
Etwas	,217	,160	,195	,046	,023	,028	,107
Nein	,507	,611	,519	,033	,019	,026	,265
Aktiver Rand	1,000	1,000	1,000	1,000	1,000	1,000	

z.B. 167/607=0,275 =27,5% der bis 39jährigen Männer können eine Hose umnähen.

- Masse der Zeilen

Das **durchschnittliche Spaltenprofil** (Durchschnittsspaltenprofil) erhält man, indem die Randsummen der Zeilen ins Verhältnis zur Gesamtsumme gesetzt werden.

-Schwerpunkt

Das durchschnittliche Spaltenprofil wird **als Vergleichsmaßstab** für die einzelnen Spaltenprofile verwendet. Es liegt immer im Mittelbereich der Spaltenprofilwerte; sie sind deren gewichtete Mittelwerte. *Das durchschnittliche Spaltenprofil ist daher* ***der*** **Schwerpunkt der Spaltendarstellung.**

Analog: Zeilenprofil

Zeilenprofile

	GeschAlt						
Hose umnähen	M 0-39	M 40-59	m 60 -	W 0-39	w 40-59	w 60 -	Aktiver Rand
Ja	,087	,057	,048	,334	,262	,213	1,000
Etwas	,405	,233	,190	,098	,037	,037	1,000
nein	,381	,360	,204	,028	,012	,014	1,000
Masse	,199	,156	,104	,227	,171	,141	

167/1910=0,0874 8,7% derjenigen, die Hosen umnähen können, sind männlich und unter 40 Jahren

Zeilenprofile, durchschnittliches Zeilenprofil, Masse der Spalte, Schwerpunkt der Zeilendarstellung.

-Überführbarkeit

Das Verhältnis von einem Zeilenprofilelement zu dem durchschnittlich Zeilenprofilelement ist gleich dem Verhältnis des entsprechenden Spaltenprofilelements zu dem durchschnittlichen Spaltenprofilelement.
M bis 39, Hose umnähen, ja > 0,2751/0,6275 = 0,0874/0,1994.

Anstatt standardisierte Profile, „gewichtete Profile“ über die Zeilen und die Spaltensummen.
0,1994*0,2751/0,6275=0,0874 = Zeilenprofil
Werte der Zeilendarstellung erscheinen als gewichtete Werte der Spaltendarstellung.

3. Chi-Quadrat-Distanz

Distanzen:
Bsp. Messung des Weges, den ein Objekt in Manhatten zurücklegt. (Auto, Hubschrauber)

City-Block-Metrik:
Punkte (A,B) Distanz = |a1-b1| + |a2 – b2|

Euklidische Distanz:

Distanz = $\sqrt{(a1 - b1)^2 + (a2 - b2)^2}$

$$\chi^2 = \Sigma\Sigma \frac{(Nij - \grave{}Nij)^2}{\grave{}Nij}$$

Beispiel: Zusammenhang Interviewbereitschaft/Umfang der Informationsschreiben für ein Interview

	Anschreibe n			
Interview	Ohne A 1	Kurzes A 2	Langes A 3	Aktiver Rand
Interviewt 1	**32**	46	36	*114*
Verweiger. 2	62	32	48	142
Nicht-ang. 3	17	19	13	49
Ausf. 4	13	20	24	57
Aktiver Rand	124	117	121	**362**

Spaltenprofile

	ANSCHR			
INTERV	1	2	3	Masse
1	,258	,393	,298	**,315**
2	,500	,274	,397	,392
3	,137	,162	,107	,135
4	,105	,171	,198	,157
Aktiver Rand	1,000	1,000	1,000	

Zeilenprofile

	ANSCHR			
INTERV	1	2	3	Aktiver Rand
1	,281	,404	,316	1,000
2	,437	,225	,338	1,000
3	,347	,388	,265	1,000
4	,228	,351	,421	1,000
Masse	**,34**3	,323	,334	

0,315*0,343=0,108 362*0,108=39,1 Erwartungswert für Zelle (1,1)
oder (114*124)/362 = 39,1 Konzept der statistischen Unabhängigkeit

$$\chi^2 = \Sigma\Sigma \frac{(Nij - \grave{}Nij)^2}{Nij} = \frac{(39,1 - 32)^2}{39,1} + \frac{(36,8-46)^2}{36,8} \ldots$$

$$\chi^2 = \frac{(39,1 - 32)^2}{\frac{114}{39,1}} + \ldots = 114 * \frac{(0,343 - \mathbf{0,281})}{0,343}$$

Kürzung durch Randsumme
Nij = empirisches Zeilenprofil (1,1)=0,281
`Nij = erwartetes Zeilenprofil =0,343 = durchschnittliches Zeilenprofil

$$\frac{\chi^2}{362} = \frac{***0,315***}{} * \frac{(0,343 - 0,281)^2}{0,343} + \ldots$$

Normierung mit Division durch n=362
(Normierung wie z.B. bei Cramers V oder Phi²
114/362 = 0,315 = durchschnittliches Spaltenprofil

Gesamtträgheitsgewicht = die quadrierte Abweichung von dem erwartete und beobachteten Profilelement, gewichtet mit den korrespondierenden durchschnittlichen Zeilen und Spaltenprofilen

Durch Ausklammerung des gemeinsamen Multiplikators können die Elemente der ersten Zeile zusammengefasst werden:

$$\frac{\chi^2}{362} = 0{,}315 * \left[\frac{(0{,}343 - 0{,}281)^2}{0{,}343} + \frac{(0{,}323 - 0{,}404)^2}{0{,}323} + \frac{(0{,}334 - 0{,}316)^2}{0{,}334} \right] + \ldots$$

$$d^*(Z,I) = \sqrt{\left[(0{,}343 - 0{,}281)^2 + (0{,}323 - 0{,}404)^2 + (0{,}334 - 0{,}316)^2 \right]}$$

Diese Distanzen (d*) entsprechen den Euklidischen Distanzen zwischen dem Schwerpunkt der Zeilendarstellung (Z) und der ersten Zeile (durchgeführte Interviews =I).

Wenn die einzelnen Summanden durch das je durchschnittliche Zeilenprofil dividiert werden, dann bezeichnet man diese Distanzen als „gewichtete Euklidische Distanzen“ oder als „**Chi-Quadrat-Distanzen**“ (d^2). Aufgrund dieser Distanzen spricht man von **Chi-Quadrat-Metrik.**

$$d^2(Z,I) = \sqrt{\frac{(0{,}343 - 0{,}281)^2}{0{,}343} + \frac{(0{,}323 - 0{,}404)^2}{0{,}323} + \frac{(0{,}334 - 0{,}316)^2}{0{,}334}} = 0{,}180$$

Mit dieser Metrik können auch Chi-Quadrat-Distanzen zwischen zwei Profilen berechnet werden. Die Distanz zwischen der ersten Zeile (Interviewte = I) und der zweiten Zeile (Verweigerer = VW) beträgt 0,414:

$$d^2(I,VW) = \sqrt{\frac{(0{,}281 - 0{,}437)^2}{0{,}343} + \frac{(0{,}404 - 0{,}225)^2}{0{,}323} + \frac{(0{,}316 - 0{,}338)^2}{0{,}334}} = 0{,}414$$

Die Distanz zwischen den Interviewten und dem Zeilendurchschnittsprofil (Schwerpunkt) ist also geringer als zwischen den Interviewten und den Verweigerern.

	Interviewter	Verweigerungen	Nicht-Angetroffen	Sonstige Ausfälle
Verweigert	0,414	-	-	-
Nicht-Angetroffen	0,146	0,349	-	-
Sonstige Ausfälle	0,223	0,444	0,344	-
Schwerpunkt	0,180	0,236	0,165	0,252

Die Herleitung der Distanzen für die Spaltenprofile erfolgt analog derjenigen der Zeilenprofile

---

Zur grafischen Darstellung der Profile

1. ungewichtete Darstellung – Projektion der Spaltenprofile
2. Gewichtung
3. der optimale Unterraum
4. 2 Darstellungsarten:

symmetrische Darstellung („französische Schule“):
Eine simultane Darstellung der drei Zeilen- und der sechs Spaltenprofile.

asymmetrische Darstellung („holländische Schule“):
Die Darstellung der Profile der Zeilen/Spalten mit den Scheitelpunkte der Spalten/Zeilen.

Literatur: Blasius, Jörg: Korrespondenzanalyse. 2001, München. Oldenbourg-Verlag