
Multivariate Verfahren – Korrespondenzanalyse

Seufert
07.11.11
FSU Jena

Korrespondenzanalyse

1. **Ziel der Korrespondenzanalyse**
 - ◆ Visualisierung von Häufigkeiten qualitativer Merkmale
2. **Kontingenz-(Kreuz-)Tabellen**
 - ◆ Zeilen- und Spaltenprofile, Massen
 - ◆ Chi-Quadrat-Statistik und Totale Inertia
3. **Standardisierung der Daten**
 - ◆ Relative Häufigkeiten und Zentrierung
4. **Extraktion der Dimensionen**
 - ◆ Singulärwertzerlegung und Eigenwertanteile
5. **Normalisierung von Koordinaten**
 - ◆ Symmetrische und asymmetrische Reskalierung
6. **SPSS-Dateneingabe**
 - ◆ Tables, Casewise, Weight

Korrespondenzanalyse – Ziel und Vorgehensweise

Korrespondenzanalyse	Gibt es einen Zusammenhang zwischen qualitativen Merkmalen mit hoher Zahl an Ausprägungen? Ziel: Visualisierung des Zusammenhangs in wenigen Dimensionen mit geringst möglichem Informationsverlust	
Anwendungsbereiche z.B.	Vokale Berufstyp Konsumtyp Marke Persönlichkeitstyp	↔ Konsonanten ↔ Parteipräferenz ↔ Markenpräferenz ↔ Eigenschaft ↔ Einstellung
Vorgehensweise	<ol style="list-style-type: none"> Kreuztabellierung der Ausgangsdaten (I x J - Kreuz-/Kontingenztabelle) Standardisierung der Ausgangsdaten (Umformen in relative Häufigkeiten und Zentrierung) Extraktion der Dimension (Zerlegung in Zeilenelemente, Singulärwerte und Spaltenelemente) Normalisierung der Koordinaten (Reskalierung zur Gewinnung von Koordinaten, die Darstellung in der gleichen Ebene ermöglichen) 	<p>Bezeichnungen Elemente</p> <p>n_{ij} A</p> <p>z_{ij} Z</p> <p>u_{ik}, s_k, v_{jk} U * S * V'</p> <p>r_{ik} U → R</p> <p>c_{jk} V' → C</p>

Korrespondenzanalyse – Unterschied zur Kontingenzanalyse

Ausgangsdaten:
Merkmalsträger/Objekte mit zwei qualitativen Merkmalen mit mehreren (nominal-skalierten) Ausprägungen

I x J - Kreuztabelle		Merkmal 2 (X)				Zeilen- bzw. Randsumme
		Ausprägung				
Merkmal 1 (Y)		1	2	...	J	
Ausprägung	1	n_{11}	n_{12}		n_{1j}	$n_{1.}$
	2	n_{21}	n_{22}		n_{2j}	$n_{2.}$
	...					
	I	n_{i1}	n_{i2}		n_{ij}	$n_{i.}$
Spalten- bzw. Randsumme		$n_{.1}$	$n_{.2}$		$n_{.j}$	n

n = Gesamtzahl der Beobachtungen
 n_{ij} = Zahl der Beobachtungswerte mit der Kombination
 a) i-te Ausprägung des ersten Merkmals
 b) j-te Ausprägung des zweiten Merkmals

Kontingenzanalyse

Gibt es einen nicht-zufälligen „statistischen“ Zusammenhang zwischen den Merkmalen?
(Hypothenprüfung)

Korrespondenzanalyse

Visualisierung des Zusammenhangs zwischen den Merkmalen in 2 Dimensionen mit minimalem Informationsverlust

Korrespondenzanalyse – Zeilen- und Spaltenprofile, Massen

Ausgangsdaten: Verteilung der Fälle auf Merkmalskombinationen n_{ij}
N = 125

6 x 3 - Kreuztabelle	Automarke			Zeilen- summe
Haushaltstyp	Marke1	Marke2	Marke3	
1-Pers: 1 Erw.	5	7	11	23
2-Pers: 2 Erw.	14	17	13	44
2-Pers: 1 Erw./1Ki.	2	6	7	15
3-Pers: 3 Erw.	9	2	6	17
3-Pers: 2 Erw./1Ki.	2	7	2	11
3-Pers: 1 Erw./2Ki.	6	4	5	15
Spaltensumme	38	43	44	125

n_{ij} / n_i n_i Zeilensumme

Zeilenprofile der HH-Typen			
0,217	0,304	0,478	1,000
0,400	0,267	0,295	1,000
0,133	0,400	0,467	1,000
0,529	0,118	0,353	1,000
0,182	0,636	0,182	1,000
0,400	0,267	0,333	1,000
0,304	0,344	0,352	1,000

Durchschnittsprofil der Zeilen

Masse der Spalte 1 (Marke1)
 $p_j = n_j / n$

Durchschnittsprofil der Spalten

Masse der Zeile 5 (2E,1K)
 $p_i = n_i / n$

n_{ij} / n_j n_j Spaltensumme

Spaltenprofile der Marken			
0,132	0,163	0,250	0,184
0,368	0,395	0,295	0,352
0,053	0,140	0,159	0,120
0,237	0,047	0,136	0,136
0,053	0,163	0,045	0,088
0,158	0,093	0,114	0,120
1,000	1,000	1,000	1,000

Korrespondenzanalyse – Chi-Quadrat-Werte und „Totale Inertia“

$\chi^2 = \text{„Chi-Quadrat“} = \sum \frac{(\text{beobachtete Häufigkeit} - \text{erwartete Häufigkeit})^2}{\text{erwartete Häufigkeit}}$

erwartete Häufigkeit
 $e_{ij} = \frac{\text{Zeilensumme} \times \text{Spaltensumme}}{\text{Gesamtsumme}}$

z.B. $e_{11} = \frac{23 \times 38}{125} = 6,992$

χ^2 -Abweichung für n_{11} $(5 - 6,992)^2 / 6,992 = 0,567515$

6 x 3 - Kreuztabelle	Automarke			Zeilen- summe
Haushaltstyp	Marke1	Marke2	Marke3	
1-Pers: 1 Erw.	5	7	11	23
2-Pers: 2 Erw.	14	17	13	44
2-Pers: 1 Erw./1Ki.	2	6	7	15
3-Pers: 3 Erw.	9	2	6	17
3-Pers: 2 Erw./1Ki.	2	7	2	11
3-Pers: 1 Erw./2Ki.	6	4	5	15
Spaltensumme	38	43	44	125

Totale Inertia („Trägheit“) einer Kreuztab.

$$T = \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{\chi^2}{n}$$

Trägheitsgewichte der Zeilen

$$T_i = \frac{1}{n} \sum_{j=1}^J \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Trägheitsgewichte der Spalten

$$T_j = \frac{1}{n} \sum_{i=1}^I \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

$$0 \leq T \leq \min\{I, J\} - 1$$

z.B. 6x3Tab T = 2
6x4 Tab T = 2
11x4 Tab T = 3

Korrespondenzanalyse – Standardisierung

absolute Häufigkeit n_{ij}

6 x 3 - Kreuztabelle Haushaltstyp	Automarke			Zeilen- summe
	Marke1	Marke2	Marke3	
1-Pers: 1 Erw.	5	7	11	23
2-Pers: 2 Erw.	14	17	13	44
2-Pers: 1 Erw./1Ki.	2	6	7	15
3-Pers: 3 Erw.	9	2	6	17
3-Pers: 2 Erw./1Ki.	2	7	2	11
3-Pers: 1 Erw./2Ki.	6	4	5	15
Spaltensumme	38	43	44	125

1.Schritt:

relative Häufigkeiten

$$p_{ij} = n_{ij} / n$$

6 x 3 - Kreuztabelle Haushaltstyp	Automarke			Zeilen- summe
	Marke1	Marke2	Marke3	
1-Pers: 1 Erw.	0,040	0,056	0,088	0,184
2-Pers: 2 Erw.	0,112	0,136	0,104	0,352
2-Pers: 1 Erw./1Ki.	0,016	0,048	0,056	0,120
3-Pers: 3 Erw.	0,072	0,016	0,048	0,136
3-Pers: 2 Erw./1Ki.	0,016	0,056	0,016	0,088
3-Pers: 1 Erw./2Ki.	0,048	0,032	0,040	0,120
Spaltensumme	0,304	0,344	0,352	1,000

2.Schritt:

Zentrierung

$$z_{ij} = \frac{(p_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

$$\hat{e}_{ij} = \text{erwartete relative Häufigkeiten}$$

$$T = \sum_i \sum_j z_{ij}^2$$

Korrespondenzanalyse – Singulärwertzerlegung

$$Z = U \cdot S \cdot V'$$

(z_{ij}) (I x J - Matrix) mit standardisierten Daten
 (u_{ik}) (I x K - Matrix) für Zeilenelemente
 (s_{kk}) (K x K - Diagonalmatrix) mit Singulärwerten
 (v_{jk}) (J x K - Matrix) für Spaltenelemente

(u _{ik})	Dim1	Dim2
1-Pers: 1 Erw.		
2-Pers: 2 Erw.		
2-Pers: 1 Erw./1Ki.		
3-Pers: 3 Erw.		
3-Pers: 2 Erw./1Ki.		
3-Pers: 1 Erw./2Ki.		

Dim1	Dim2
s_{k1}^2	0
0	s_{k2}^2

gleich "Eigenwerte" der Dimension

(v _{jk})	Dim1	Dim2
Marke 1		
Marke 2		
Marke 3		

$$T = \sum_i \sum_j z_{ij}^2 = \sum_i T_i = \sum_k s_k^2 = \sum_j T_j$$

Welcher Anteil an der Gesamtstreuung wird von der k-ten Dimension aufgenommen?

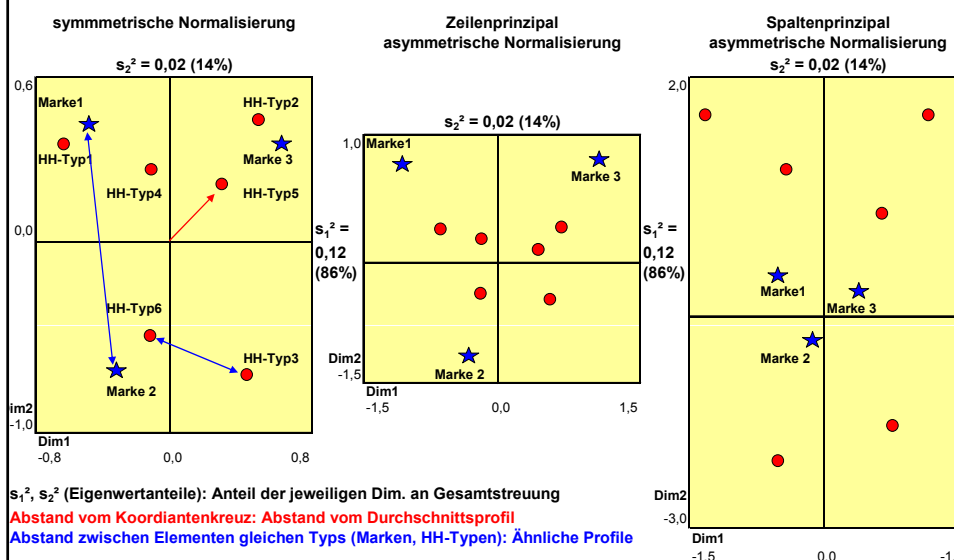
$$EA_k = \frac{s_k^2}{T} \text{ Eigenwertanteil der Dimension k}$$

Korrespondenzanalyse – (a)symmetrische Normalisierung

Ziel: Koordinaten für beide Merkmale in die gleiche Ebene legen!

Normalisierung	Zeilenpunkte (u_{ik}) → r_{ik}	Spaltenpunkte (v_{jk}) → c_{jk}
Symmetrisch	$u_{ik} * \frac{\sqrt{s_k}}{\sqrt{p_i}}$	$v_{jk} * \frac{\sqrt{s_k}}{\sqrt{p_j}}$
Zeilen-Prinzipal (asymmetrisch)	$u_{ik} * \frac{s_k}{\sqrt{p_i}}$	$v_{jk} * \frac{1}{\sqrt{p_j}}$
Spalten-Prinzipal (asymmetrisch)	$u_{ik} * \frac{1}{\sqrt{p_i}}$	$v_{jk} * \frac{s_k}{\sqrt{p_j}}$
Prinzipal (symmetrisch)	$u_{ik} * \frac{s_k}{\sqrt{p_i}}$	$v_{jk} * \frac{s_k}{\sqrt{p_j}}$
Carroll/III/Green/Schaffer (symmetrisch)	$u_{ik} * \frac{\sqrt{1 + s_k}}{\sqrt{p_i}}$	$v_{jk} * \frac{\sqrt{1 + s_k}}{\sqrt{p_j}}$

Korrespondenzanalyse – Interpretation der Konfigurationen



Korrespondenzanalyse – SPSS-Dateneingabe

Problem:

Kreuztabellen entsprechen nicht der SPSS - Dateistruktur (Fälle / Variablen)

Weg 1:

"Tables"-Eingabeformat

generierte Variable

Haushaltstyp	Marke1	Marke2	Marke3
1	5	7	11
2	14	17	13
3	2	6	7
4	9	2	6
5	2	7	2
6	6	4	5

Weg 2:

"Casewise"-Eingabeformat

HH-Typ Markenpräf

HH-Typ	Markenpräf
Fall1	1 5
Fall2	2 5
Fall3	2 2
Fall4	1 2
Fall5	2 1
...	

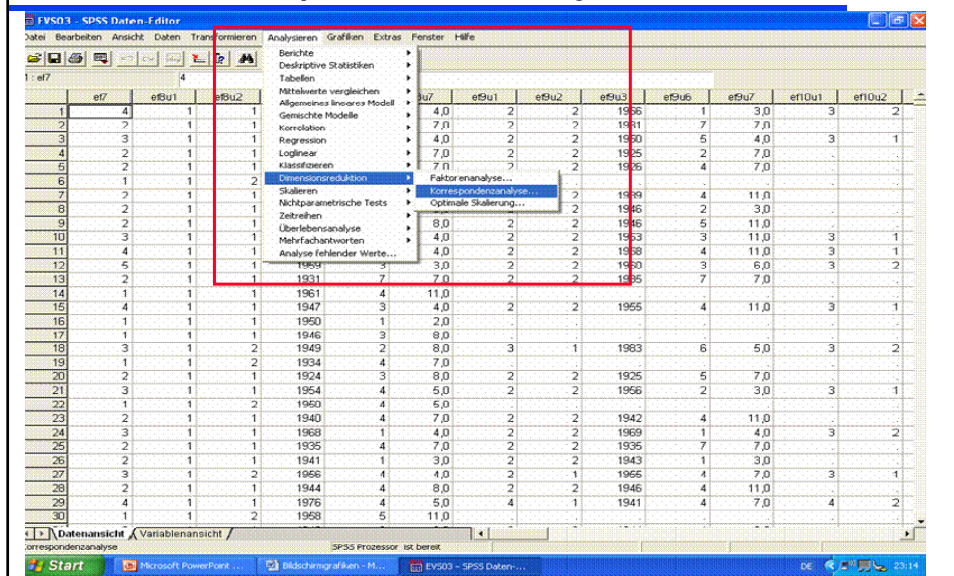
Weg 3:

"Weight"-Eingabeformat

HH-Typ Markenpräf Fallzahl

HH-Typ	Markenpräf	Fallzahl
1	1	5
1	2	7
1	3	11
2	1	14
2	2	17
2	3	13

SPSS-Beispiel Aufruf Korrespondenzanalyse



SPSS-Beispiel Vorgaben

The image displays five overlapping dialog boxes for SPSS Correspondence Analysis:

- Korrespondenzanalyse:** The main dialog box. The 'Zelle:' field contains 'elSu1? ?'. The 'Spalte:' field contains 'elSu2? ?'. Buttons for 'Bereich definieren...', 'Abbrechen', 'Hilfe', 'Modell...', 'Statistiken...', and 'Diagramme...' are visible.
- Korrespondenzanalyse: Spaltenbereich definieren:** Shows 'Kategorienbereich für Spaltenvariable: elSu2' with 'Minimalwert: 1' and 'Maximalwert: 6'. An 'Aktualisieren' button is present. Below, 'Nebenbedingungen für Kategorien' are listed (1-6) with radio buttons for 'Keine', 'Kategorien müssen gleich sein', and 'Ergänzende Kategorie'.
- Korrespondenzanalyse: Modell:** 'Dimensionen in der Lösung:' is set to 2. 'Distanzmaß' has 'Chi-Quadrat' selected. 'Standardisierungsmethode' has 'Zeilen- und Spaltenmittel werden entfernt' selected.
- Korrespondenzanalyse: Statistiken:** 'Korrespondenztabelle', 'Übersicht der Zeilenpunkte', and 'Übersicht der Spaltenpunkte' are checked. 'Permutationen der Korrespondenztabelle' is unchecked.
- Korrespondenzanalyse: Diagramme:** 'Steuertdiagramme' has 'Biplot' checked. 'Linendiagramme' has 'Transformierte Zeilenkategorien' and 'Transformierte Spaltenkategorien' unchecked.

Red arrows indicate the flow of information: from the main dialog to the column range dialog, then to the model dialog, and finally to the statistics and diagrams dialogs.