

Training Fiche Vorlage

Titel	Korrespondenzanalyse (Correspondence Analysis)	
Schlüsselwörter (Meta-Tags)	Korrespondenzanalyse, qualitative Variablen, erklärte Trägheit, Eigenwerte	
Sprache	Deutsch	
Zielsetzungen / Lernziele / Lernergebnisse	<p>Ziel dieses Moduls ist es, die Technik der Korrespondenzanalyse vorzustellen und zu erklären.</p> <p>In diesem Modul zeigen wir dir:</p> <ul style="list-style-type: none"> - Wie die Logik einer Korrespondenzanalyse funktioniert - Welche Anforderungen eine Korrespondenzanalyse hat - Wie du eine Korrespondenzanalyse selbst durchführen kannst - Wie du das FactoMineR-Paket in R für Korrespondenzanalysen verwenden kannst 	
Lehrgang:		
Datenwissenschaftliche Kompetenz		
Modul Datenvisualisierung und visuelle Analyse	X	
Einführung in die Datenwissenschaft für Human- und Sozialwissenschaften		
Datenwissenschaft für den guten Zweck		
Datenjournalismus und Geschichtenerzählen		
Beschreibung	<p>In diesem Schulungsmodul lernen wir die multidimensionale Analysetechnik der Korrespondenzanalyse kennen.</p> <p>Die Korrespondenzanalyse ist eine Form der multidimensionalen Skalierung, bei der im Wesentlichen eine Art räumliches Modell erstellt wird, das die Assoziationen zwischen einer Reihe von kategorialen Variablen aufzeigt. Umfasst der Satz nur zwei Variablen, wird die Methode gewöhnlich als einfache Korrespondenzanalyse (Simple Correspondence Analysis) bezeichnet.</p>	



	<p>Umfasst die Analyse mehr als zwei Variablen, wird sie in der Regel als Multiple Korrespondenzanalyse (Multiple Correspondence Analysis) bezeichnet. Ziel dieser Analyse ist es, die Dimensionalität des untersuchten Phänomens zu reduzieren und gleichzeitig die darin enthaltenen Informationen zu erhalten. Diese Methode kann nur mit qualitativen Variablen verwendet werden.</p> <p>Der letzte Teil des Moduls ist der Anwendung der Korrespondenzanalyse (CA) in der Software R gewidmet.</p>
<p>Inhalt in 3 Ebenen gegliedert</p>	<p>1. EINLEITUNG</p> <p>Die Korrespondenzanalyse ist eine mehrdimensionale Analysetechnik, mit der fast jede Art von Tabelle, die aus numerischen Daten besteht, in grafische Form übertragen werden kann. Gegenstand der Korrespondenzanalyse sind Kontingenzmatrizen, deren Elemente angeben, wie oft die Merkmale zweier unterschiedlicher Größen zusammen festgestellt wurden.</p> <p>Das Hauptziel der Korrespondenzanalyse ist die Analyse der Beziehungen zwischen zwei variablen und qualitativen Merkmalen, die an einem Kollektiv von statistischen Einheiten beobachtet werden. Dies geschieht durch die Identifizierung eines "optimalen" Raums, d. h. einer reduzierten Dimension, die die Synthese der in den ursprünglichen Daten enthaltenen Strukturinformationen darstellt. Ziel der Analyse ist es, die zwischen den untersuchten Daten bestehenden Verflechtungen oder Korrespondenzen aufzuzeigen.</p> <p>2. ANFORDERUNGEN AN DIE KORRESPONDENZANALYSE</p> <p>Um eine Korrespondenzanalyse durchführen zu können, ist es wichtig, die zu verwendenden Variablen zu analysieren, um einige ihrer Eigenschaften zu klären. Insbesondere müssen die Variablen die folgenden Anforderungen erfüllen:</p> <ul style="list-style-type: none"> - <i>Die Variablen müssen qualitativ sein:</i> Qualitative Variablen sind Variablen, die nicht durch Zahlen, sondern durch Modalitäten dargestellt werden, z. B.: Geschlecht, Bildungsniveau, Familienstand usw. Diese Modalitäten, auch Kategorien genannt, müssen <u>erschöpfend</u> sein und <u>sich gegenseitig ausschließen</u>. <u>Gegenseitiger Ausschluss</u> bedeutet, dass die variablen Modalitäten nicht dieselbe Art von Information enthalten dürfen. Zum Beispiel können wir für die Variable "Haarfarbe" nicht die Modalitäten "dunkles Haar" und "braunes Haar" angeben, da dunkles Haar auch braunes Haar bedeutet und umgekehrt. <u>Erschöpfend</u> bedeutet, dass die Modalitäten einer Variable alle Möglichkeiten berücksichtigen müssen. Zum Beispiel werden für die Variable "Bildungsgrad"



die Modalitäten "Diplom", "Bachelor", "Hochschulabschluss" eingefügt. Diese drei Modalitäten berücksichtigen nicht alle möglichen Bildungsabschlüsse.

- *Die Variablen müssen voneinander abhängig sein:*

Vor der Durchführung der Korrespondenzanalyse muss der Grad der gegenseitigen Abhängigkeit zwischen den beiden betrachteten Variablen überprüft werden, da eine Analyse der Übereinstimmungen nicht sinnvoll wäre, wenn sie unabhängig wären.

Zu diesem Zweck führen wir den Chi-Quadrat-Test durch:

H_0 : die beiden Variablen sind unabhängig

H_1 : die beiden Variablen sind nicht unabhängig

Um die Ergebnisse des Tests zu interpretieren, betrachten wir den p-Wert:

p-Wert < 0,05: Die Nullhypothese wird abgelehnt, und folglich wird davon ausgegangen, dass die Variablen einen gewissen Grad an Abhängigkeit aufweisen.

3. Wie man eine Korrespondenzanalyse durchführt

Nachdem wir die -Anforderungen überprüft haben, können Sie mit der eigentlichen Analyse beginnen.

3.1) Kontingenztabelle

In der Korrespondenzanalyse arbeiten wir mit Kontingenztabelle, welche die gemeinsamen Häufigkeiten der Werte der beiden qualitativen Variablen X und Y enthalten. Diese Matrizen bestehen immer aus ganzen Zahlen, welche nie negativ sein können. Diese ganzen Zahlen sind Zählungen, d. h. einfache Aufzeichnungen des Auftretens. Außerdem spielen die beiden kategorialen Variablen eine symmetrische Rolle, bei der alle Elemente denselben Charakter haben.

$X \setminus Y$	y_1	y_2	y_3	
x_1				
x_2		$n_{i,j}$		n_i
x_3				
		n_j		n

X, Y sind die qualitativen Variablen.

x_1, x_2, x_3 : sind die Werte von X



y_1, y_2, y_3 : sind die Werte von Y

$n_{i,j}$: sind die absoluten gemeinsamen Häufigkeiten, d. h. die Häufigkeiten der Paare, z. B. $n_{1,1}$: $X = x_1; Y = y_1$

$n_{i.}$: sind die Zeilenränder: $n_{i.} = \sum_{j=1}^C n_{i,j}$

$n_{.j}$: sind die Spaltenränder: $n_{.j} = \sum_{i=1}^R n_{i,j}$

Dies ist die Summe der gemeinsamen Häufigkeiten der Werte von Y (für die Spalten der Werte von X) für die feste Zeile (oder Spalte).

n = die Anzahl der Stichproben, welche durch Addition der Ränder der Zeilen oder Spalten ermittelt werden kann:

$$n = \sum_{i=1}^R \sum_{j=1}^C n_{i,j} \quad \forall i, j$$

Wir können von absoluten Frequenzen zu relativen Frequenzen wechseln, indem wir jede absolute Frequenz durch n dividieren:

$$f_{i,j} = \frac{n_{i,j}}{n}$$

3.2) Zeilenprofilmatrix und Spaltenprofilmatrix

Die Zeilenprofilmatrix erhält man, indem man die absoluten Häufigkeiten (oder relativen Häufigkeiten) durch die jeweiligen Zeilenränder dividiert. Daher:

$$\frac{n_{i,j}}{n_{i.}} = \frac{f_{i,j}}{f_{i.}} \quad \forall i, j$$

Die Kontingenztabelle wird wie folgt aussehen:

		1
	$\frac{f_{i,j}}{f_{i.}} = \frac{n_{i,j}}{n_{i.}}$	1
		1
	profilo medio	1

An den Rändern der Reihen haben wir alle 1, und dies entspricht der Summe der Reihenprofile.

An den Spaltenrändern befinden sich die Durchschnittsprofile, die durch Addition der relativen Häufigkeiten pro Spalte oder durch Mittelwertbildung der Elemente



der Zeilenprofilmatrix pro Spalte ermittelt werden. Es handelt sich um einen gewichteten Durchschnitt, wobei die Massen durch die Zeilenränder f_i dargestellt werden.

Bei der Arbeit mit Frequenzen geht eine Dimension verloren, so dass der Zeilenraum durch einen Raum mit C-1 Dimensionen dargestellt wird.

Das bedeutet, dass eine **diagonale Matrix von Zeilenrandhäufigkeiten D_R** konstruiert werden kann, die auf der Hauptdiagonalen Zeilenprofile aufweist. Die Diagonalmatrix der **Zeilenrandhäufigkeiten** ist eine Matrix **R·R**, deren Dimensionen der Anzahl der Zeilen entspricht und die auf der Hauptdiagonale die Zeilenränder der relativen Häufigkeitstabelle enthält. Eine Diagonalmatrix ist eine Matrix, dessen Werte nur auf der Hauptdiagonale nicht Null sind (darüber oder darunter gleichen alle Werte Null). Sie ist immer eine symmetrische und quadratische Matrix. Mit der Diagonalmatrix der **Zeilenrandhäufigkeiten** kann man die **Reihe der Zeilenprofilmatrixen** konstruieren: Man erhält sie, indem man die relativen Häufigkeiten durch die Zeilenränder dividiert $\frac{F}{D_R}$. Die Dimensionen von **F** sind R·C, während **D_R** die Dimension R·R hat. Da die Division zwischen Matrizen nicht möglich ist, berechnet man den Kehrwert von **D_R** und multipliziert mit **F**, wodurch das Dimensionalitätsproblem gelöst wird: **$D_R^{-1} \cdot F$** .

Das Gleiche gilt für die Spalten, mit einigen kleinen Unterschieden.

Die Spaltenprofilmatrix wird erstellt, indem die absoluten Häufigkeiten durch die jeweiligen Spaltenrandhäufigkeiten dividiert werden:

$$\frac{n_{i,j}}{n_j} = \frac{f_{i,j}}{f_j} \quad \forall i,j$$

Dadurch erhalten wir folgende Kontingenztable:

				profilo
	$\frac{f_{i,j}}{f_j} = \frac{n_{i,j}}{n_j}$			medio
1	1	1	1	1

In diesem Fall gleichen an den Rändern der Spalte alle Werte 1 und an den Rändern der Zeile erhalten wir das durchschnittliche Spaltenprofil. In diesem Fall werden die



Massen durch die Spaltenränder $f_{.j}$ dargestellt. Offensichtlich arbeitet man auch im Spaltenraum mit weniger als einer Dimension, also ist der Spaltenraum R-1.

Es kann eine **diagonale Matrix von Spaltenrandhäufigkeiten** D_C konstruiert werden, die Spaltenprofile auf der Hauptdiagonale hat. Die diagonale Matrix der **Spaltenrandhäufigkeiten** ist eine Matrix $C \cdot C$, deren Dimensionen den Spalten entsprechen und die auf der Hauptdiagonale die Spaltenränder der relativen Häufigkeitstabelle enthält. Mit der Diagonalmatrix der Spaltenränder kann man die **Matrix der Spaltenprofile** konstruieren: Man erhält sie, indem man die relativen Häufigkeiten durch die Spaltenränder dividiert $\frac{F}{D_R}$. Die Dimensionen von F sind $R \cdot C$, während D_C die Dimension $C \cdot C$ hat. Da die Division zwischen den Matrizen nicht möglich ist, berechnet man den Kehrwert von D_C und multipliziert sie mit F , wodurch das Problem der Dimensionalität gelöst wird:

$$F \cdot D_C^{-1}.$$

3.3) Distanzen berechnen

Bei der Korrespondenzanalyse ist es notwendig zu verstehen, welcher Abstand zwischen den Werten besteht, um zu verstehen, ob die Modalitäten weit oder nahe beieinander liegen und ob sie sich daher ähneln oder nicht. Dies ist möglich, indem man die Frequenzen betrachtet: je niedriger sie sind, desto näher sind sie sich und umgekehrt. Es gibt verschiedene Methoden zur Berechnung des Abstands:

Euklidischer Abstand und Chi-Quadrat-Abstand.

Der **euklidische Abstand** ist der einfachste und identifiziert die größten Abstände auf Kosten der kleinsten. Er wird berechnet, indem die Differenz der relativen Häufigkeiten zum Quadrat erhoben wird.

Für Reihenprofile:

$$d_{(i,v)} = \sqrt{\sum_{j=1}^C \left(\frac{f_{i,j}}{f_{i.}} - \frac{f'_{i',j}}{f'_{i'.}} \right)^2}$$

Für Spaltenprofile:



$$d_{(j,j')} = \sqrt{\sum_{i=1}^R \left(\frac{f_{i,j}}{f_{\cdot j}} - \frac{f_{i,j'}}{f_{\cdot j'}} \right)^2}$$

Der **Chi-Quadrat-Test** identifiziert kleinere Abstände. Hier werden die Häufigkeiten mit geringer Anzahl in Bezug auf die Zeilen neu gewichtet, indem der Kehrwert der Spaltenrandhäufigkeiten in die Formel eingefügt wird (in Bezug auf die Spalten wird der Kehrwert der Zeilenrandhäufigkeiten in die Formel eingefügt). Der Nachteil des Chi-Quadrat-Tests besteht darin, dass der Kehrwert der Spalten- (oder Zeilen-) Randhäufigkeiten gegen Null tendieren kann und daher eine einzelne Antwort übermäßig zur Berechnung des Abstands beitragen kann.

3.4) Zeilenraum und Spaltenraum

Im **Zeilenraum** gibt es zwei Komponenten:

- Reihenprofil: $\mathbf{D}_R^{-1} \cdot \mathbf{F}$
- Metrisch: \mathbf{D}_C^{-1}

Beginnen wir mit der Formel:

$$\boldsymbol{\Psi}_{n \times 1} = \mathbf{X}_{n \times p} \cdot \mathbf{u}_{p \times 1}$$

Indem Sie geeignete Ersetzungen vornehmen:

$$\boldsymbol{\Psi} = \mathbf{D}_R^{-1} \cdot \mathbf{F} \cdot \mathbf{D}_C^{-1} \cdot \mathbf{u}$$

Das Ziel der Korrespondenzanalyse ist die Menge der Einheitsachsen, die es ermöglichen, die Abstände zwischen den Projektionen der Reihenprofile zu maximieren. Wir müssen also nach den Vektoren suchen, die die Projektionen maximieren. Da Vektoren \mathbf{u} unendlich sein können, wird die Einheitsnorm-Beschränkung hinzugefügt.

$$\mathbf{u}^T \cdot \mathbf{D}_C^{-1} \cdot \mathbf{u} = 1$$

Maximierungsproblem: Maximierung der erklärten Trägheit (erklärte Variation), die der Variabilität für quantitative Variablen entspricht.

$$\left\{ \begin{array}{l} \text{MAX: } \{ \hat{\boldsymbol{\psi}}^T \mathbf{D}_R \hat{\boldsymbol{\psi}} \} \\ \mathbf{v}^T \mathbf{D}_C^{-1} \mathbf{v} = 1 \end{array} \right.$$



Um das Problem der eingeschränkten Maximierung zu lösen, verwenden Sie die Methode der Lagrange-Multiplikatoren:

$$\mathcal{L}(v, \lambda) = (\hat{\psi}^T D_R \hat{\psi}) - \lambda(v^T D_C^{-1} v - 1)$$

λ = Lagrange-Multiplikator, der ein Skalar ist;

u = Vektor der gesuchten Gewichte

Wenn wir die notwendigen Ersetzungen vornehmen, haben wir:

$$\mathcal{L}(v, \lambda) = (D_R^{-1} F D_C^{-1} v)^T D_R (D_R^{-1} F D_C^{-1} v) - \lambda(v^T D_C^{-1} v - 1)$$

Wir führen die Transpositionsoperationen durch, ersetzen $D_R \cdot D_R^{-1}$ für die Identitätsmatrix I und $[(-\lambda) \cdot (-1)]$ ersetzen sie durch λ . Wir können dann die Transponierung aus den Diagonalmatrizen D_C^{-1} und D_R^{-1} entfernen, da sich die Transponierung einer Diagonalmatrix nicht ändert. Erhalten:

$$\mathcal{L}(v, \lambda) = v^T D_C^{-1} F^T D_R^{-1} F D_C^{-1} v - \lambda v^T D_C^{-1} v + \lambda$$

Wir berechnen die partiellen Ableitungen, leiten die Lagrange ab in Bezug auf u und setzen sie gleich 0:

$$\frac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow \frac{\partial \mathcal{L}}{\partial v} = 2F^T D_R^{-1} F D_C^{-1} v - 2\lambda v = 0$$

Multiplizieren Sie die Gleichung mit D_C^{-1} :

$$F^T D_R^{-1} F D_C^{-1} v = \lambda v$$

Wenn wir die Transponierung der Zeilenprofile und die Matrix der Spaltenprofile durch S ersetzen, können wir die charakteristische Gleichung wie folgt schreiben:

$$Sv = \lambda v$$



Die Maximierung der erklärten Trägheit von Zeilenprofilen ist gleichbedeutend mit der Zerlegung dieser Matrix in Eigenwerte und Eigenvektoren derselben. Der erste Eigenwert ist mit dem ersten Eigenvektor verbunden, der die maximale Trägheit erklärt. Die Eigenvektoren, die anschließend extrahiert werden, werden orthogonal extrahiert.

$$\mathbf{u}_1^T \cdot \mathbf{D}_C^{-1} \cdot \mathbf{u}_2 = 0$$

Wir verwenden die Orthogonalitätsbeschränkung, um die zweite Komponente auszuwählen, die die Trägheit erklärt, die nicht von der ersten Komponente erklärt wird. Offensichtlich erklärt die erste extrahierte Komponente die maximale Trägheit, d. h. die maximale Dehnung der Punktwolke.

In den **Spalten Raum** sind zwei Komponenten:

- Spaltenprofil: $\mathbf{F} \cdot \mathbf{D}_C^{-1}$
- Metrisch: \mathbf{D}_R^{-1}

Beginnen wir mit der Formel:

$$\boldsymbol{\varphi}_{p \times 1} = (\mathbf{X}_{n \times p}^T)_{p \times n} \cdot \mathbf{v}_{n \times 1}$$

Wir ersetzen und erhalten

$$\boldsymbol{\varphi} = \mathbf{D}_C^{-1} \mathbf{F}^T \mathbf{D}_R^{-1} \mathbf{v}$$

Das mit Lagrange-Multiplikatoren zu lösende Maximierungsproblem lautet:

$$\begin{cases} \text{MAX: } \{\hat{\boldsymbol{\varphi}}^T \mathbf{D}_C \hat{\boldsymbol{\varphi}}\} \\ \mathbf{v}^T \mathbf{D}_R^{-1} \mathbf{v} = 1 \end{cases}$$

Wenn wir wie im Raum der Zeilen vorgehen, erhalten wir schließlich:



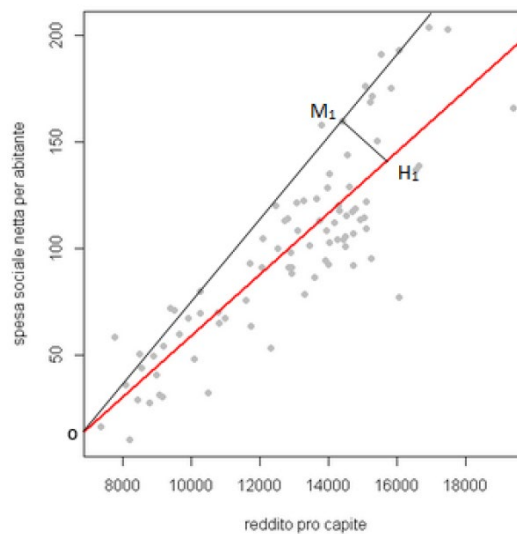
$$FD_C^{-1}F^T D_R^{-1}\nu = \mu\nu$$

Setzt man die Matrix der Spaltenprofile und die transponierte Metrik der Zeilenprofile mit S^* erhält man die charakteristische Gleichung:

$$S^*\nu = \mu\nu$$

Die geometrische Maximierung der erklärten Trägheit, d.h. die verlorene Information so klein wie möglich und die beobachtete Information so groß wie möglich zu machen, lautet: den Abstand M_1H_1 so klein wie möglich und die Entfernung OH_1 so groß wie möglich.

Figura 1.3: Diagramma di dispersione



Wir müssen also die Gerade f (in rot) finden, die die Punkte des Vektorraums so interpoliert, dass der Abstand zwischen allen Punkten des Raums und den orthogonal auf die Gerade f projizierten Punkten so gering wie möglich ist.

Die Eigenwerte im Zeilenraum entsprechen den Eigenvektoren im Spaltenraum, d. h. die Eigenwerte von S entsprechen den Eigenwerten von S^* . Die Eigenvektoren sind bis auf eine Konstante einander gleich. Wenn wir also maximieren müssen, brauchen wir nicht in Eigenwerte und Eigenvektoren zu zerlegen S und S^* sondern nur in einen.



Der Betrag der erklärten Trägheit ist gleich, ob wir \mathbf{S} oder \mathbf{S}^* berechnen, wird die Beziehung zwischen den beiden Räumen durch die **Übergangsformeln** dargestellt:

$$\mathbf{S} \rightarrow \boldsymbol{\nu} = \frac{1}{\sqrt{\lambda}} \mathbf{F} \mathbf{D}_C^{-1} \boldsymbol{\nu} \equiv \mathbf{S}^* \rightarrow \boldsymbol{\nu} = \frac{1}{\sqrt{\lambda}} \mathbf{F}' \mathbf{D}_R^{-1} \boldsymbol{\nu}$$

Reihen Platz:

$$\hat{\boldsymbol{\psi}} = \mathbf{D}_C^{-1} \boldsymbol{\nu}$$

Mit:

$$\boldsymbol{\nu} = \frac{1}{\sqrt{\lambda}} \mathbf{F}' \mathbf{D}_R^{-1} \boldsymbol{\nu}$$

Durch Anwendung der entsprechenden Substitutionen:

$$\frac{1}{\sqrt{\lambda}} \mathbf{D}_C^{-1} \mathbf{F}' \mathbf{D}_R^{-1} \boldsymbol{\nu} \rightarrow \frac{1}{\sqrt{\lambda}} \mathbf{D}_C^{-1} \mathbf{F}' \hat{\boldsymbol{\psi}}$$

Erhalten:

$$\sqrt{\lambda} \hat{\boldsymbol{\psi}} = \mathbf{D}_C^{-1} \mathbf{F}' \hat{\boldsymbol{\psi}} \rightarrow \hat{\boldsymbol{\psi}} = \frac{1}{\sqrt{\lambda}} \mathbf{D}_C^{-1} \mathbf{F}' \hat{\boldsymbol{\psi}} \rightarrow \sqrt{\lambda} \hat{\boldsymbol{\psi}} = \mathbf{D}_C^{-1} \mathbf{F}' \hat{\boldsymbol{\psi}}$$

Für den Raum der Zeilen, daher:

$$\sqrt{\lambda} \hat{\boldsymbol{\psi}} = \mathbf{D}_C^{-1} \mathbf{F}' \hat{\boldsymbol{\psi}} \equiv \hat{\boldsymbol{\psi}} = \sqrt{\lambda} \hat{\boldsymbol{\psi}}$$

Spalten Platz:

$$\hat{\boldsymbol{\psi}} = \mathbf{D}_R^{-1} \boldsymbol{\nu}$$

Wo:



$$\nu = \frac{1}{\sqrt{\lambda}} F D_C^{-1} v$$

Durch Anwendung der entsprechenden Substitutionen:

$$\frac{1}{\sqrt{\lambda}} D_R^{-1} F D_C^{-1} v \rightarrow \frac{1}{\sqrt{\lambda}} D_R^{-1} F \hat{\psi}$$

Erhalten:

$$\frac{1}{\sqrt{\lambda}} D_R^{-1} F \hat{\psi} \rightarrow \sqrt{\lambda} \hat{\psi} \rightarrow D_R^{-1} F \hat{\psi}$$

Für den Spaltenraum:

$$\sqrt{\lambda} \hat{\psi} = D_R^{-1} F \hat{\psi} \equiv \hat{\psi} = \sqrt{\lambda} \hat{\psi}$$

4) Beispiel mit R-Software

Überprüfung eines möglichen Zusammenhangs zwischen der Verteilung des Viehbestands und den verschiedenen italienischen Regionen. Die Daten beziehen sich auf das Jahr 2011 und wurden von den auf der Istat-Website verfügbaren Banken erhoben.

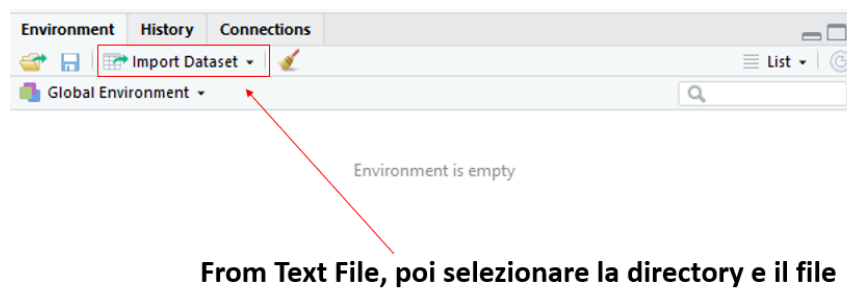
Hypothese: Die verschiedenen Regionen entscheiden sich je nach den territorialen Merkmalen und den Bedürfnissen der Bevölkerung dafür, bestimmte Rinder zu züchten und andere nicht.

Datensatz:



Regione	Bovini	Ovini	Caprini	Equini	Suini	Conigli	Totale
Piemonte	23516	2303	3418	2370	2429	1392	35428
Valle d'Aosta	1585	347	284	53	16	11	2296
Liguria	1642	1126	549	949	258	924	5448
Lombardia	15480	2592	3175	3647	4346	1191	30431
Trentino Alto Adige	10482	2279	2424	1513	3292	266	20256
Veneto	16007	1642	1207	2429	3634	1907	26826
Friuli-Venezia Giulia	1539	83	207	280	1477	117	3703
Emilia-Romagna	8522	1315	908	3161	1541	308	15755
Toscana	4392	4918	607	2163	2046	1764	15890
Umbria	3132	2734	667	1245	4107	1924	13809
Marche	2940	1877	342	383	7103	1786	14431
Lazio	9256	8678	1624	3535	6849	4269	34211
Abruzzo	5588	6590	1710	1362	10241	2450	27941
Molise	2976	2510	610	534	3943	60	10633
Campania	10971	6248	3675	1448	15145	6708	44195
Puglia	3010	1918	826	691	759	921	8125
Basilicata	3156	7426	3562	1280	6137	2606	24167
Calabria	5496	3701	3505	1839	21522	2087	38150
Sicilia	7387	4963	1088	1930	821	63	16252
Sardegna	8200	12880	3171	3333	9324	523	37431
Totale	145277	76130	33559	34145	104990	31277	425378

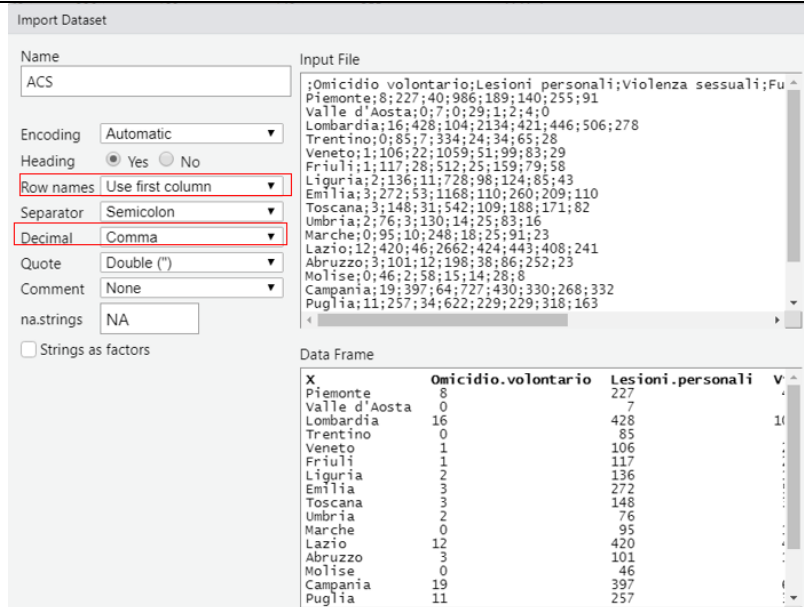
Wir importieren den Datensatz:



Wählen Sie im Feld **row names** die Formulierung: "**use first column**", um die Bezeichnungen der Individuen und Variablen in den Diagrammen zu haben.

Im Dezimalfeld wählen wir "**comma**".





X	Omicidio.volontario	Lesioni.personali	V
Piemonte	8	227	7
Valle d'Aosta	0	7	10
Lombardia	16	428	10
Trentino	0	85	10
Veneto	1	106	10
Friuli	1	117	10
Liguria	2	136	10
Emilia	3	272	10
Toscana	3	148	10
Umbria	2	76	10
Marche	0	95	10
Lazio	12	420	10
Abruzzo	3	101	10
Molise	0	46	10
Campania	19	397	10
Puglia	11	257	10

Mit dem Befehl:

X<-as.matrix(nome_del_dataset)

Wir ordnen **X** als Objekt den in der Analyse verwendeten Datensatz zu.

Vor der Durchführung des Korrespondenzanalyse ist es notwendig, den Grad der gegenseitigen Abhängigkeit zwischen den beiden betrachteten Charakteren festzustellen, denn wenn sie unabhängig sind, macht es möglicherweise keinen Sinn, das Korrespondenzanalyse fortzusetzen. Um dies zu überprüfen, führen wir den Chi-Quadrat-Test durch.

Der Befehl lautet:

chiquadro<-chisq.test(X)

Pearson's Chi-squared test

data: X

X-squared = 126691.2, df = 95, p-value < 2.2e-16

Es ist festzustellen, dass der **p-Wert** niedriger ist als das am häufigsten verwendete Signifikanzniveau, d. h. 0,05. Wir können daher die Nullhypothese der statistischen Unabhängigkeit zwischen den beiden Variablen verwerfen und mit der Analyse fortfahren.



Nun wollen wir eine Matrix der relativen Häufigkeiten **F** erstellen.

Wir berechnen die Stichprobenzahl mit dem Befehl:

```
n<-sum(X)
```

und dividieren dann die Ausgangsmatrix (also alle gemeinsamen Häufigkeiten) durch die Stichprobenzahl, um die Matrix **F** zu erhalten. Dazu verwenden wir folgenden Befehl:

```
F<-X/n
```

Der nächste Schritt besteht darin, die **Zeilen- und Spaltenprofil**tabellen zu erhalten. Dazu ist es zunächst erforderlich, die Randhäufigkeiten der Zeilen und Spalten zu berechnen. Die entsprechenden Befehle lauten:

```
sumrow<-apply(F,1,sum)  
sumcol<-apply(F,2,sum)
```

Dann berechnen wir mit den Befehlen die Diagonalmatrix der Randhäufigkeiten der Zeile und ihre Umkehrung:

```
Dr<-diag(sumrow)  
Dr_inv<-solve(Dr)
```

Nun können wir die Zeilenprofile berechnen. In Matrixform multiplizieren wir die Umkehrung der Randreihen-Diagonalmatrix mit der Matrix der relativen Häufigkeiten. Dazu verwenden wir folgenden Befehl:

```
Pr<-Dr_inv%*%F
```

Das Gleiche gilt für Spaltenprofile, wobei zu beachten ist, dass in diesem Fall die Inverse der Spaltenmatrix mit der Matrix der relativen Häufigkeiten nachmultipliziert werden muss.

```
Dc<-diag(sumcol)  
Dc_inv<-solve(Dc)  
Pc<-F%*%Dc_inv
```

Nun können wir die Abstände zwischen den Punkten berechnen. Wie bereits erwähnt, gibt es zwei Arten von Abständen: **Euklidisch** und **Chi-Quadrat**.

Euklidischer Abstand der **Reihenprofile**:



```
d_euc_r<-dist(rbind(Pr[1,],Pr[2,]))
```

Euklidischer Abstand der **Spaltenprofile**:

```
d_euc_c<-dist(rbind(Pr[,1],Pr[,2]))
```

Abstand der **Chi-Quadrat-Reihenprofile**:

```
d_r<-pr[1,]-pr[2,]
d<-d_r^2/sumcol
d_chi_r<-sqrt(Summe(d))
```

Abstand der **Chi-Quadrat-Spaltenprofile**:

```
dc<-Pr[,1]-Pr[,2]
dc<-dc^2/sumrow
d_chi_c<-sqrt(Summe(dc))
```

Die charakteristische Gleichung der Zeilenprofilmatrix:

```
S<-t(Pr)%*%Pc
```

Da die Matrix S nicht symmetrisch ist, muss sie diagonalisiert werden, um **S_tilde** zu erhalten:

```
A<-t(F)%*%Dr_inv%*%F #Symmetrie
```

```
Dc_12<-diag(sumcol^(-1/2))
```

```
S_tilde<-Dc_12%*%A%*%Dc_12
```

Nun gilt es, die durch Zerlegung der Matrix in Eigenwerte und Eigenvektoren erklärte Trägheit zu maximieren:

```
AC<-eigen(S_tilde)
```

```
lambda<-as.matrix(AC$Werte)
```

```
lambda<-lambda[-1,]
```

```
w<-AC$Vektoren
```




```
u<-Dc^(1/2)%*%w
```

```
u<-u[,-1]
```

Die charakteristische Gleichung der Spaltenprofilmatrix :

```
S_star<-F%*%Dc_inv%*%t(F)%*%Dr_inv
```

Um sich von u nach v zu bewegen, verwenden wir Übergangsformeln (da der Betrag der erklärten Trägheit sowohl im Zeilen- als auch im Spaltenraum gleich ist).

```
sq_lambda<-diag((sqrt(lambda))^-1)
```

```
v<-F%*%Dc_inv%*%u%*%sq_lambda
```

Wir berechnen Faktoren und Koordinaten, zuerst den Zeilenraum und dann die Spalten:

```
fp_r<-Dc_inv%*%u
```

```
fp_c<-Dr_inv%*%v
```

```
PHI_coord<-Dc_inv%*%t(F)%*%fp_c
```

```
PSI_coord<-Dr_inv%*%F%*%fp_r
```

Wir zeigen die Grafik der Hauptkoordinaten an:

```
PRINCOORD<-rbind(PSI_coord,PHI_coord)
```

```
rows<-row.names(X);columns<-colnames(X)
```

```
plot(PRINCOORD[,1],PRINCOORD[,2],type="n",main="Hauptkoordinaten",xlab="Axis1",ylab="Axis2")+
```

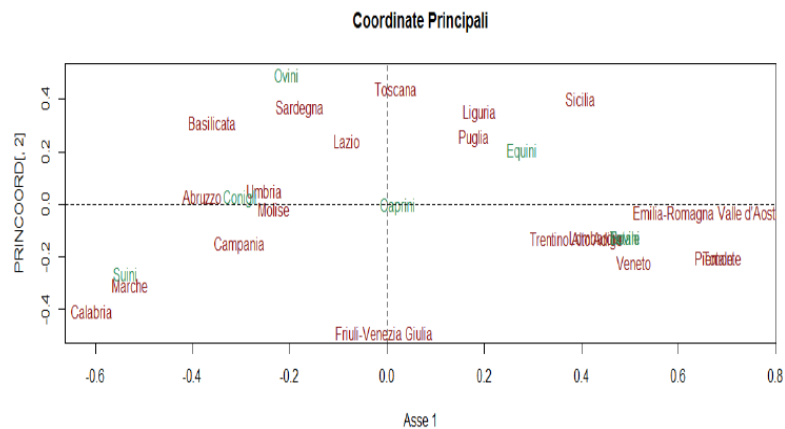
```
text(PRINCOORD[1:20,1],PRINCOORD[1:20,2],labels=rows,col="springgreen4")
```

```
text(PRINCOORD[21:29,1],PRINCOORD[21:29,2],labels=columns,col="violetred")
```

```
abline(h=0,v=0,lty=2,lwd=1.5)
```

Wir erhalten:





Anhand dieses Diagramms können wir beispielsweise feststellen, dass in Regionen wie den Abruzzen, Molise und Umbrien hauptsächlich Kaninchen gezüchtet werden.

Wir wählen die Komponenten aus:

```
inertia<-sum(diag(S))-1
```

```
sum(lambda)
```

```
in_exp<-lambda/inertia
```

```
in_exp_<-cumsum(in_exp)
```

Dann visualisieren wir die erzielten Ergebnisse:

```
> inerzia
[1] 0.2978321
> in_exp
[1] 0.58571295 0.23305781 0.10382933 0.04875445 0.02864546
> in_exp_cum
[1] 0.5857130 0.8187708 0.9226001 0.9713545 1.0000000
```

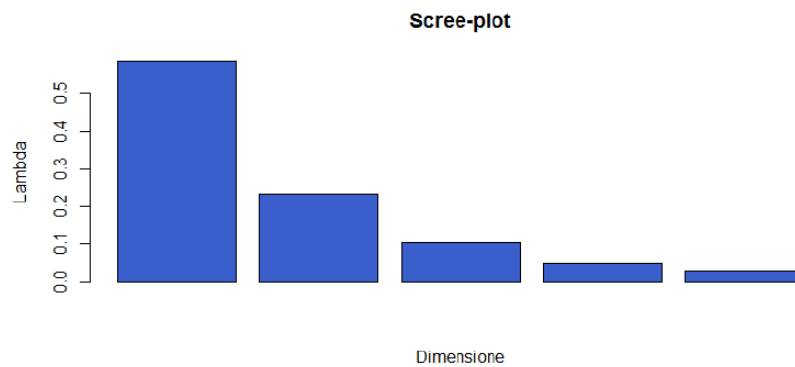
Die erste Dimension allein erklärt 58,57 % der Variabilität, und die ersten drei Dimensionen zusammen erklären 92,26 % der Gesamtvariabilität der Daten.

Die erhaltenen Ergebnisse können durch einen **Scree-plot der erklärten Trägheit** grafisch dargestellt werden:

```
screeplot<-barplot(in_exp,main="Scree-plot trägheit", xlab="Größe",
ylab="Lambda", col="hellblau")
```



Figura 1.10: Scree-plot dell'inerzia spiegata



Für die Qualität der Darstellung:

- Um zu bewerten, wie stark ein Modus die faktorielle Achse beeinflusst oder an ihr teilnimmt, berechnen wir **die absoluten Beiträge (ca)** sowohl für die Zeilen als auch für die Spalten:

```
ca_r<-Dr%%fp_c^2
```

```
ca_c<-DC%%fp_r^2
```

- Um die Qualität der Darstellung zu bewerten, berechnen wir die **relativen Beiträge, CR**. Sie sind ein besseres Maß für die Darstellung der Punkte auf den Achsen und werden durch den Kosinus des Winkels zwischen dem Projektionsvektor des Punktes und dem relativen Vektor i (oder j) an dem Punkt i (oder j) in seinem ursprünglichen Raum bestimmt:

```
G<-matrix(sumcol,20,9,byrow=T)
```

```
di<-(Pr-G)^2%%Dc_inv
```

```
d_ig<-apply(di,1,sum)
```

```
cos2r<-PSI_Koordinate^2/d_ig
```

```
H<-matrix(sumrow,20,9)
```

```
dj<-Dr_inv%%(Pc-H)^2
```

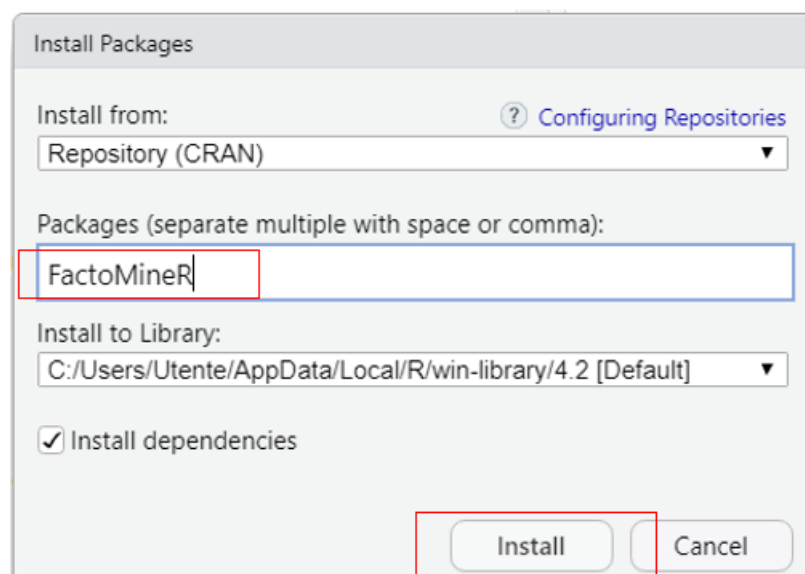
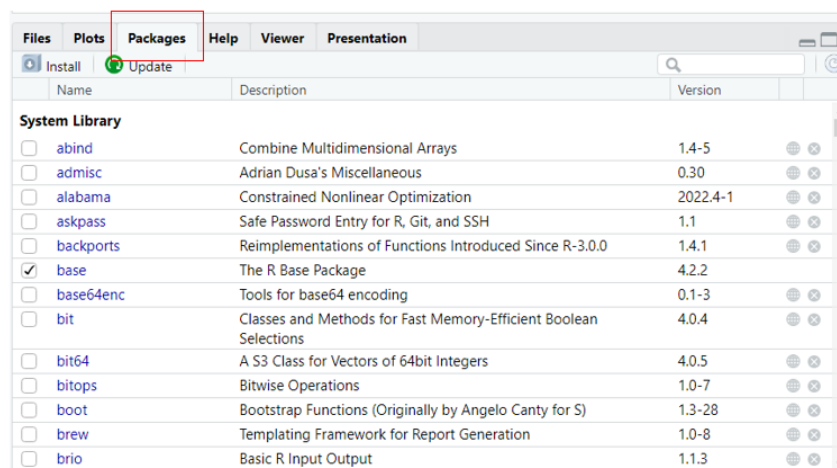
```
d_jh<-apply(dj,2,sum)
```

```
COS2C<-PHI_Koordinate^2/d_jh
```



R bietet für die Korrespondenzanalyse ein Paket namens **FactoMineR** an, das Informationen über Personen und Variablen hinzufügt und es uns ermöglicht, ein gemeinsames zweidimensionales Diagramm von Personen und Variablen zu erstellen.

Um dieses Paket in R verwenden zu können, müssen wir es zunächst herunterladen:



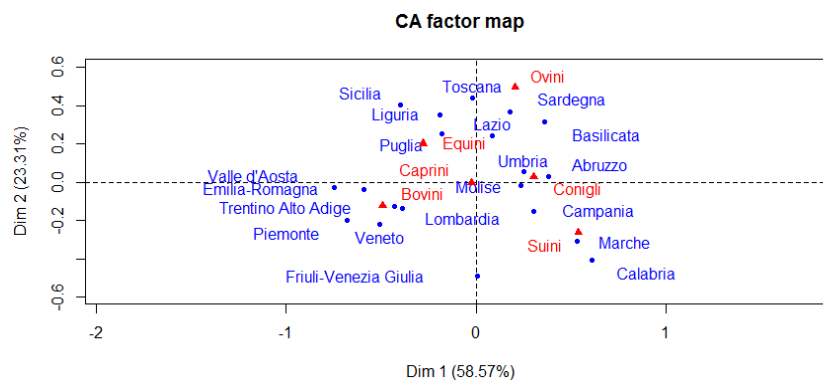
Nach der Installation führen wir das Paket mit dem folgenden Befehl aus:

Bibliothek(FactoMineR)

Gehen wir nun zur Erstellung des zweidimensionalen Graphen Individuen und Variablen über:

CA(X, ncp = 5, row.sup = NULL, col.sup = NULL, quanti.sup=NULL, quali.sup = NULL, graph = TRUE, axes = c(1,2), row.w = NULL)

Grafisch werden wir folgendes Ergebnis erhalten:



Interpretation der Ergebnisse:

Wir können sagen, dass sich die ursprüngliche Hypothese bestätigt hat. Insbesondere scheinen die Toskana, Sardinien und die Basilikata die Regionen zu sein, in denen die Schafzucht am stärksten ausgeprägt ist, was darauf zurückzuführen ist, dass es sich bei diesen Regionen um Berg- und Transhumanzgebiete handelt. Pferde werden vor allem in Apulien, Ligurien und Sizilien gezüchtet, da diese Tiere seit jeher für die Arbeit auf dem Lande eingesetzt werden. Rinder gibt es in Trentino-Südtirol, Venetien, Piemont, der Lombardei und der Emilia-Romagna; in diesen Regionen ist die Zucht für die Verwendung als Nahrungsmittel traditionell weiterentwickelt. Kaninchen kommen vor allem in Umbrien, den Abruzzen und Molise vor. Schweine werden dagegen eher in den Marken, Kampanien und Molise gezüchtet; diese Regionen haben auch eine traditionell stärker entwickelte Zucht für die Verwendung als Nahrungsmittel.

Ziegen hingegen befinden sich in der Mitte der Achsen, wahrscheinlich weil es keine Regionen gibt, in denen sie bevorzugt gezüchtet werden.

Selbstbeurteilung (Multiple-Choice)

1. Die Korrespondenzanalyse arbeitet mit:



Choice-Fragen und Antworten)	<p>A) Kontingenztabelle B) Korrelationstabelle C) Einfachen Einsätzen</p> <p>2. Warum wird der Chi-Quadrat-Test vor der Korrespondenzanalyse durchgeführt?</p> <p>A) Um zu prüfen, ob die Variablen quantitativ sind B) Um zu beurteilen, ob die Variablen qualitativ sind C) Analyse der Interdependenz zwischen den beiden Variablen</p> <p>3. Was ist das Ziel der Korrespondenzanalyse?</p> <p>A) Maximierung der erklärten Variabilität B) Maximierung der erklärten Trägheit C) Minimierung der erklärten Trägheit</p>
Ressourcen (Videos, Verweislinks)	
Verwandtes Material	
Verwandte PPT	
Literaturverzeichnis	<p>van der Heijden, P. G. M. & de Leeuw, J. (1985). Korrespondenz Analyse ergänzend zur loglinearen Analyse verwendet, Psychometrika, 50, S. 429-447.</p> <p>Le, S., Josse, J. & Husson, F. (2008). FactoMineR: Ein R-Paket für multivariate Analysen. Zeitschrift für statistische Software. 25(1). pp. 1-18.</p> <p>Mineo, A. M. (2003). Una Guida all'utilizzo dell'Ambiente Statistico R, http://cran.r-project.org/doc/contrib/Mineo-dispensaR.pdf.</p>
Zur Verfügung gestellt von	[Unisalento]

