# 4. Clusteranalyse

## 4.1 Einleitung

Die Clusteranalyse wird eingesetzt, um **Objekte** – Kunden, Regionen etc. – in Gruppen (=**Cluster**) einzuteilen. In der Marktforschung werden beispielsweise Marktsegmente mit einer Clusteranalyse definiert.

Mit der **Clusteranalyse** werden Objekte (=statistische Einheiten; Untersuchungseinheiten) in Gruppen eingeteilt, die in sich möglichst ähnlich (homogen) und untereinander möglichst verschieden (inhomogen) sein sollen.

**Beispiel 4.1**: In der Marktforschung werden Kunden mit Hilfe einer Clusteranalyse in Marktsegmente eingeteilt. Die einzelnen Marktsegmente enthalten Kunden, die relativ ähnlich, was den Bedarf, die Vorstellung vom Produkt, die Kaufhäufigkeit etc. betrifft, sind. Die Marktsegmente stellen bei genügend trennscharfer Abgrenzung unterschiedliche Zielgruppen dar, die z.B. Nachfrager ganz bestimmter Produkte sind oder für die. spezielle Werbestrategien entwickelt werden können. ♦

Im einfachsten Fall könnte man statistische Einheiten in Gruppen unter Verwendung einer Variablen durch eine **Klassifizierung** einteilen. Diese Vorgehensweise weist jedoch mehrere Nachteile auf:

- Die Klassengrenzen oder die Gruppenzahl müssen weitgehend subjektiv festgelegt werden.
- Zwei Objekte können unterschiedlichen Clustern angehören, obwohl ihre Variablenwerte relativ dicht beieinander liegen. Ein Objekt liegt dann knapp unterhalb und ein Objekt dicht oberhalb der Klassengrenze.
- Es wird nur eine Variable zur Gruppenbildung herangezogen. Sofern die Werte für mehrere Variablen bekannt sind, werden nicht alle Informationen genutzt.

**Beispiel 4.2:** Zwölf Regionen sollen unter Verwendung der Einwohnerdichte (ED) in drei Gruppen eingeteilt werden. Unter Verwendung der Klassengrenzen:

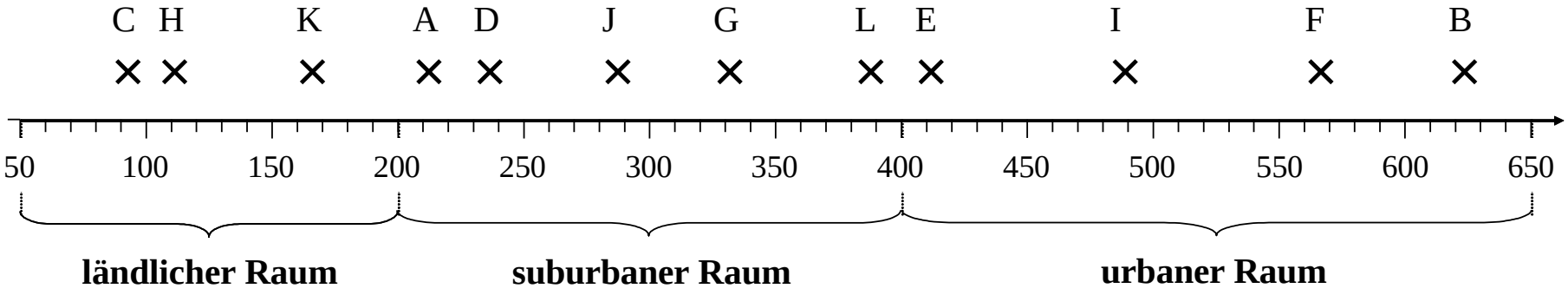
- $50 \leq ED < 200$ : ländlicher Raum
- $200 \leq ED < 400$: suburbaner Raum
- $400 \leq ED < 650$: verstädterter Raum

erhält man mit den Merkmalswerten:

| Region | Einwohnerdichte (ED) | Gruppe |
|---|---|---|
| A | 212,4 | suburbaner Raum |
| B | 623,7 | verstädterter Raum |
| C | 93,1 | ländlicher Raum |
| D | 236,8 | suburbaner Raum |
| E | 412,0 | verstädterter Raum |
| F | 566,7 | verstädterter Raum |
| G | 331,9 | suburbaner Raum |
| H | 111,4 | ländlicher Raum |
| I | 489,0 | verstädterter Raum |
| J | 287,4 | suburbaner Raum |
| K | 166,2 | ländlicher Raum |
| L | 388,1 | suburbaner Raum |

folgende Gruppen:

- Ländlicher Raum: C, H, K
- Suburbaner Raum: A, D, J, G, L
- Verstädterter (urbaner) Raum: B, E, F, I.




Von den genannten Nachteilen sollen einige aufgegriffen werden:

- Die Obergrenze der ersten Klasse hätte nicht nur 200, sondern auch 250 oder 300 lauten können.
- Die Regionen L und E haben beispielsweise eine ähnlich hohe Einwohnerdichte, sie werden aber zwei unterschiedlichen Räumen zugeordnet.
- Wir haben noch weitere Informationen über die Variablen, wie beispielsweise den Wanderungssaldo, das BIP je Einwohner etc., die zur Gruppenbildung nicht genutzt werden.

♦

Die **Clusteranalyse** setzt genau an den Schwächen des einfachen Klassifikations-verfahrens an. Sie orientiert sich bei der Problemlösung an **zwei Arbeitsschritte**.

Im 1. Schritt muss die **Ähnlichkeit bzw. Distanz** zwischen den Objekten gemessen werden.

Der 2. Schritt beinhaltet die **Gruppierung der Objekte** aufgrund ihrer Ähnlichkeiten bzw. Distanzen. Ohne eine Vorabkenntnis der Gruppenzahl (Clusterzahl) – was den Regelfall darstellt – wird hierzu ein **hierarchisches Verfahren** gewählt. Eine hiermit erhaltene Gruppierung, die nicht als optimal angesehen wird, kann mit einem **Partitionsverfahren** optimiert werden. Auf diese Weise lässt sich eine hierarchische Cluster-lösung durch Anwendung eines Partitionsverfahrens verbessern.

Die Arbeitsschritte bei der **Klassifikation von Objekten** gehen aus der folgenden Übersicht hervor:

| | | | | |
|---|---|---|---|---|
| Arbeitsschritt 1 | **Quantifizierung der Ähnlichkeit von Objektpaaren** | | | |
| | **Wahl eines Ähnlichkeitsmaßes in Abhängigkeit vom Skalenniveau** | | | |
| | **Metrische Skala** | **Ordinalskala** | **Nominalskala** | **Gemischtes Skalenniveau** |

$\Downarrow$

| | | | |
|---|---|---|---|
| Arbeitsschritt 2 | **Gruppierung der Objekte aufgrund ihrer Ähnlichkeit** | | |
| | **Wahl eines Klassifikationsverfahrens** | | |
| | Arbeitsschritt 2a | $\Rightarrow$ | Arbeitsschritt 2b |
| | **Gruppenzahl nicht bekannt** | | **Gruppenzahl bekannt** |
| | **Wahl und Einsatz eines hierarchischen Verfahrens** | | **Wahl und Einsatz eines Partitionsverfahrens** |

# 4.2 Distanz und Ähnlichkeitsmaße

## 4.2.1 Distanzmaße

### Euklidische Distanz

Zweidimensionaler Zahlenraum:
Berechnung der **direkten Distanz** zwischen zwei Punkten nach dem Satz von Pythagoras als Hypotenuse eines "gedachten" rechtwinkligen Dreiecks

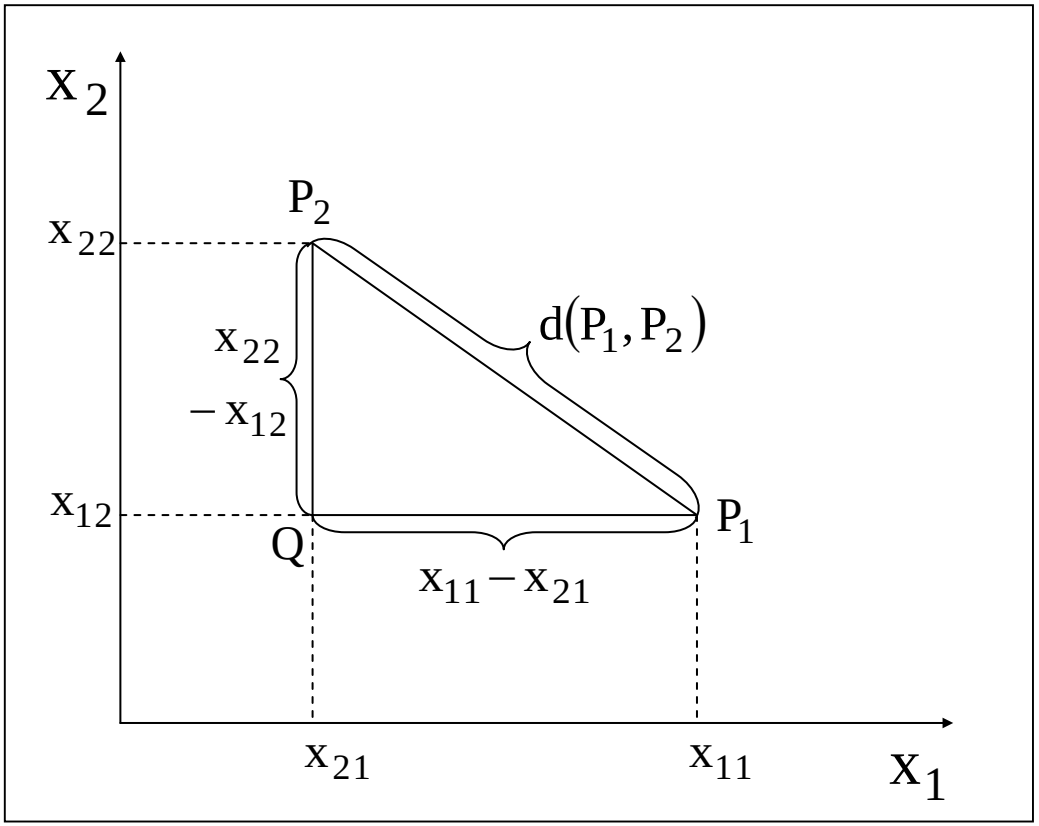
Direkte Distanz $d_{12}$ zwischen den Punkten $P_1$ und $P_2$:

Wurzel der Summe der beiden Kathetenquadrate $(x_{11} - x_{21})^2$ und $(x_{22} - x_{12})^2$

(4.1) $$d(P_1, P_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{22} - x_{12})^2}$$

Die **direkte Distanz** zwischen den beiden Punkten $P_1$ und $P_2$ heißt **euklidische Distanz**. Sie lässt sich ohne weiteres auf m Dimensionen verallgemeinern. Es handelt sich hierbei um eine intuitiv naheliegende Abstandsmessung in einem reellen Zahlenraum, die aber keinesfalls die einzig mögliche ist.

Abbildung 4.1: Euklidische Distanz in der Ebene



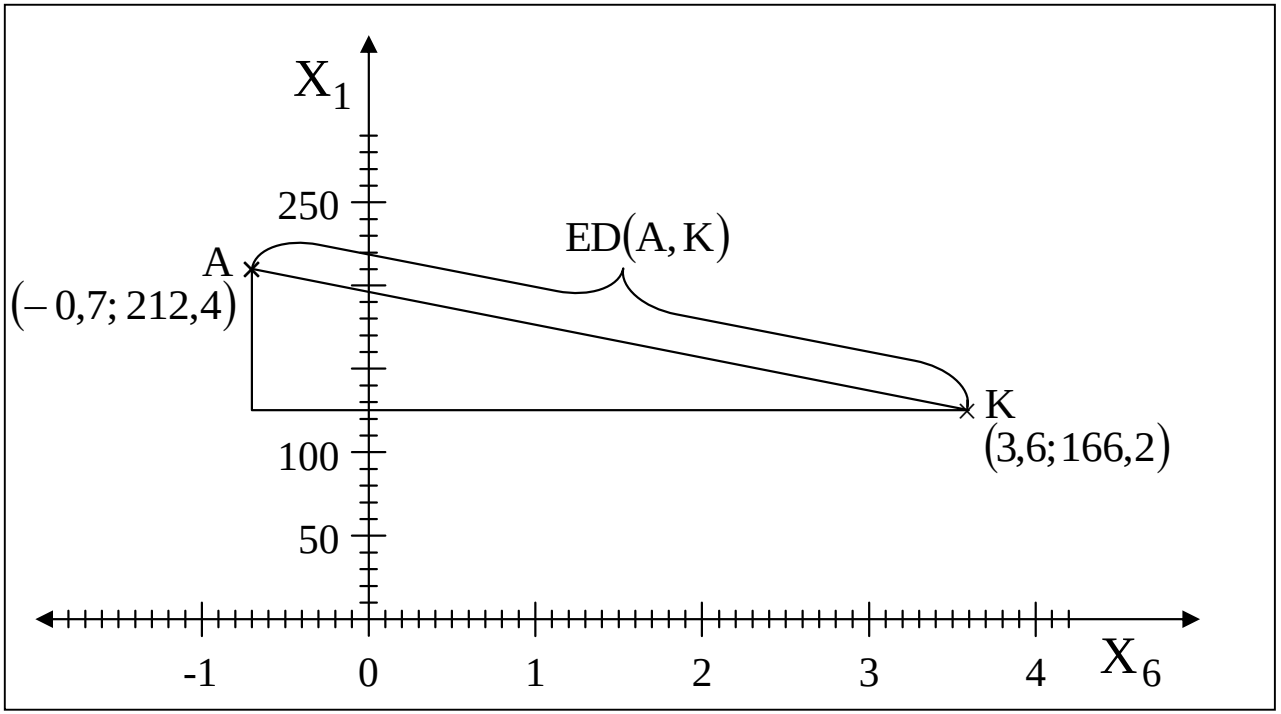

Sofern die Eigenschaften zu klassifizierender Objekte durch **metrisch skalierte Merkmale** erfasst werden können, bietet es sich unmittelbar an, ihre Ähnlichkeit unter Verwendung des Konzepts der euklidischen Distanz zu messen. Verallgemeinert man (4.1) auf einen m-dimensionalen Merkmalsraum, dann ist die **euklidische Distanz** zwischen den beiden Objekten i und j durch

(4.2) $ED(i,j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$ definiert.

**Beispiel 4.3:** Beispielhaft soll die euklidische Distanz zwischen den Regionen A und K auf der Basis der beiden Merkmale Bevölkerungsdichte ($x_1$) und Wanderungssaldo ($x_6$) bestimmt werden. Aus den aus der nachfolgenden Abbildung hervorgehenden Koordinaten nimmt die euklidische Distanz zwischen den Regionen A und K den Wert

$$
\begin{aligned}
ED(A,K) &= \sqrt{(x_{A1}-x_{K1})^2+(x_{A6}-x_{K6})^2} \\
&= \sqrt{(212{,}4-166{,}2)^2+[(-0{,}7)-3{,}6]^2} \\
&= \sqrt{46{,}2^2+(-4{,}3)^2} = \sqrt{2116+18{,}49} \\
&= \sqrt{2134{,}49} = 46{,}201
\end{aligned}
$$

Abbildung 4.2: Euklidische Distanz zwischen den Regionen A und K



Problematik:

**Ermittelte Distanz** wird durch die **Maßeinheit** bestimmt, in der die Merkmale gemessen werden. In Abhängigkeit von der Größenordnung der Merkmalswerte kann die Distanz zwischen zwei Objekten stark durch einzelne Merkmale geprägt werden. Auf diese Weise werden die **Merkmalswerte implizit ungleich gewichtet**.

Im Beispiel wird die euklidische Distanz zwischen den Regionen A und K eindeutig durch die Bevölkerungsdichte ($x_1$) dominiert. Die Distanz zwischen A und K verändert sich bei einer Hinzunahme des Wanderungssaldos ($x_6$) nur unwesentlich von 46,0 auf 46,201, obwohl die eindimensional betrachteten Werte dieses Merkmals in den beiden Regionen A und K beträchtlich differieren.

**Gleichgewichtung der Merkmale**:
Distanzmessung auf der Grundlage standardisierter Merkmalswerte
z-Transformation:

(4.3) $z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$ , i = 1, 2, ..., n; k = 1, 2, ..., m.

Die **euklidische Distanz** der **standardisierten Merkmalswerte** ist durch

(4.4) $ED(i,j) = \sqrt{\sum_{k=1}^{m} (z_{ik} - z_{jk})^2}$

gegeben.

**Beispiel 4.4:** Unter Anwendung der z-Transformation erhält man für die beiden Regionen A und K folgende Merkmalswerte für die Merkmale Bevölkerungsdichte ($x_1$) und Wanderungssaldo ($x_6$):

| Region | Bevölkerungsdichte ($x_1$) | Wanderungssaldo ($x_6$) |
|---|---|---|
| A | $z_{A1} = \frac{x_{A1} - \bar{x}_1}{s_1} = \frac{212{,}4 - 326{,}558}{173{,}885} = -0{,}657$ | $z_{A6} = \frac{x_{A6} - \bar{x}_6}{s_6} = \frac{-0{,}7 - 0{,}742}{3{,}069} = -0{,}470$ |
| K | $z_{K1} = \frac{x_{K1} - \bar{x}_1}{s_1} = \frac{166{,}2 - 326{,}558}{173{,}885} = -0{,}922$ | $z_{K6} = \frac{x_{K6} - \bar{x}_6}{s_6} = \frac{3{,}6 - 0{,}742}{3{,}069} = 0{,}931$ |

Berechnet man mit den **standardisierten Werten** der beiden Merkmalen $X_1$ und $X_6$ die **euklidische Distanz** zwischen den Regionen A und K, erhält man

$$\begin{aligned}
\mathrm{ED}(\mathrm{A},\mathrm{K}) &= \sqrt{(z_{A1} - z_{K1})^2 + (z_{A6} - z_{K6})^2} \\
&= \sqrt{[(-0{,}657) - (-0{,}922)]^2 + [(-0{,}470) - 0{,}931]^2} \\
&= \sqrt{0{,}226^2 + (-1{,}401)^2} = \sqrt{0{,}071 + 1{,}963} \\
&= \sqrt{2{,}034} = 1{,}426\,.
\end{aligned}$$

Bei Verwendung der standardisierten Merkmalswerte zeigt sich, dass die Regionen A und K hinsichtlich des Wanderungssaldos erheblich unähnlicher sind als in Bezug auf die Bevölkerungsdichte. Im Folgenden bestimmen wir die Objektdistanzen stets auf der Basis der standardisierten Merkmalswerte.

## Standardisierung mit SPSS

Eine Standardisierung von Daten lässt sich mit SPSS durch Wahl der Menüpunkte

Analysieren
  Deskriptive Statistiken
    Deskriptive Statistiken…

durchführen. Hierzu sind die zu standardisierenden Variablen im Fenster „Deskriptive Statistik“ zu markieren und durch Anklicken der Pfeilschaltfläche in das Feld „Variable(n):“ zu übertragen. Anschließend ist das Item „Standardisierte Werte als Variablen speichern“ mit einem Häkchen zu versehen.

In unserem Fall übertragen wir die sechs quantitativen Variablen der Datei Regionen(12).sav in das Feld „Variable(n):“ SPSS berechnet dann die standardisierten Merkmalswerte aller sechs Variablen und gibt sie unter Verwendung des Präfixes „Z“ im SPSS-Dateneditor aus:

Regionen - SPSS Daten-Editor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafiken Extras Fenster Hilfe

1 : ed 212,4

| | region | Zed | Zbip | Zel | Zwbip | Zgeb | Zws |
|---|---|---|---|---|---|---|---|
| 1 | A | -,657 | -1,245 | -,073 | -,450 | -,450 | -,470 |
| 2 | B | 1,709 | 1,254 | -1,242 | 1,542 | -1,306 | ,866 |
| 3 | C | -1,343 | -1,653 | 2,448 | -,955 | 1,002 | -,861 |
| 4 | D | -,516 | ,299 | -,274 | ,917 | -,338 | ,410 |
| 5 | E | ,491 | ,280 | -,238 | -1,055 | -,599 | -1,252 |
| 6 | F | 1,381 | 1,022 | -,749 | -1,312 | -,375 | -1,219 |
| 7 | G | ,031 | -,178 | -,512 | ,006 | ,257 | 1,290 |
| 8 | H | -1,237 | -,988 | 1,114 | ,620 | 1,597 | 1,453 |
| 9 | I | ,934 | 1,275 | -,822 | -,807 | -1,082 | -1,089 |
| 10 | J | -,225 | ,311 | -,256 | ,184 | 1,039 | ,312 |
| 11 | K | -,922 | -,940 | ,713 | 1,631 | 1,262 | ,932 |
| 12 | L | ,354 | ,563 | -,110 | -,321 | -1,008 | -,372 |

Datenansicht Variablenansicht

SPSS Prozessor ist bereit

Hinweis: Die nicht-standadisierten Variablen bleiben bei einer Standardisierung erhalten. Sie sind nur der besseren Übersicht halber in der obigen Datenansicht des SPSS-Dateneditors herausgenommen worden. ♦

# Berechnung der euklidischen Distanzen mit SPSS

Die **Berechnung der euklidischen Distanzen** zwischen allen 12 Regionen lässt sich mit **SPSS** durch Wahl der Menüpunkte

Analysieren
    Korrelation
        Distanzen…

durchführen.
Im Fenster „Distanzen“ übertragen wir hierzu die standardisierten Variablen der Regionen(12).sav-Datei in das Feld „Variablen“. Die euklidische Distanz braucht nicht extra gewählt zu werden, da sie bereits ebenso wie das generelle Maß „Unähnlichkeit“ voreingestellt ist.

**Näherungsmatrix**

| | Euklidisches Distanzmaß | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1:A | 2:B | 3:C | 4:D | 5:E | 6:F | 7:G | 8:H | 9:I | 10:J | 11:K | 12:L |
| 1:A | .000 | 4.438 | 3.084 | 2.259 | 2.161 | 3.326 | 2.368 | 3.293 | 3.217 | 2.423 | 3.163 | 2.151 |
| 2:B | 4.438 | .000 | 6.777 | 2.887 | 3.895 | 3.710 | 3.222 | 5.374 | 3.188 | 3.640 | 4.708 | 2.948 |
| 3:C | 3.084 | 6.777 | .000 | 4.339 | 4.128 | 5.191 | 4.351 | 3.228 | 5.373 | 3.885 | 3.697 | 4.360 |
| 4:D | 2.259 | 2.887 | 4.339 | .000 | 2.781 | 3.461 | 1.595 | 3.004 | 3.022 | 1.590 | 2.453 | 1.857 |
| 5:E | 2.161 | 3.895 | 4.128 | 2.781 | .000 | 1.312 | 2.969 | 4.623 | 1.360 | 2.679 | 4.453 | 1.263 |
| 6:F | 3.326 | 3.710 | 5.191 | 3.461 | 1.312 | .000 | 3.428 | 5.398 | 1.020 | 3.149 | 5.221 | 1.943 |
| 7:G | 2.368 | 3.222 | 4.351 | 1.595 | 2.969 | 3.428 | .000 | 2.666 | 3.337 | 1.403 | 2.601 | 2.299 |
| 8:H | 3.293 | 5.374 | 3.228 | 3.004 | 4.623 | 5.398 | 2.666 | .000 | 5.410 | 2.529 | 1.292 | 4.176 |
| 9:I | 3.217 | 3.188 | 5.373 | 3.022 | 1.360 | 1.020 | 3.337 | 5.410 | .000 | 3.168 | 5.122 | 1.452 |
| 10:J | 2.423 | 3.640 | 3.885 | 1.590 | 2.679 | 3.149 | 1.403 | 2.529 | 3.168 | .000 | 2.349 | 2.309 |
| 11:K | 3.163 | 4.708 | 3.697 | 2.453 | 4.453 | 5.221 | 2.601 | 1.292 | 5.122 | 2.349 | .000 | 3.902 |
| 12:L | 2.151 | 2.948 | 4.360 | 1.857 | 1.263 | 1.943 | 2.299 | 4.176 | 1.452 | 2.309 | 3.902 | .000 |

Dies ist eine Unähnlichkeitsmatrix

Man erhält damit eine **Distanzmatrix**, die

- symmetrisch ist, da es gleich ist, ob z.B. die Distanz zwischen A und K oder K und A gemessen wird,
- deren Diagonalelemente gleich 0 sind, da sie die Eigendistanzen der Regionen darstellen.

Mit einem Wert der euklidischen Distanz von 1,020 erweisen sich die Regionen F und I auf der Basis der ausgewählten Merkmale als am ähnlichsten. Die beiden unähnlichsten Regionen sind dagegen B und C mit einer euklidischen Distanz von 6,777. ♦

Die Berechnung der euklidischen Distanz mit allen 6 Merkmalen sei beispielhaft für die Regionen A und K nachvollzogen:

$$
\begin{aligned}
ED(A,K) &= \sqrt{\sum_{k=1}^{6}\left(z_{A,k} - z_{K,k}\right)^2} \\
&= \Big\{\left[(-0{,}657)-(-0{,}922)\right]^2 + \left[(-1{,}245)-(-0{,}940)\right]^2 + \left[(-0{,}073)-0{,}713\right]^2 \\
&\quad + \left[(-0{,}450)-1{,}631\right]^2 + \left[(-0{,}450)-1{,}262\right]^2 + \left[(-0{,}470)-0{,}931\right]^2\Big\}^{0{,}5} \\
&= \left[0{,}226^2 + (-0{,}305)^2 + (-0{,}786)^2 + (-2{,}081)^2 + (-1{,}712)^2 + (-1{,}401)^2\right]^{0{,}5} \\
&= \sqrt{10{,}005} = 3{,}163 \ .
\end{aligned}
$$

## Quadrierte euklidische Distanz

Vor allem Varianz-orientierte Klassifikationsverfahren machen von den **quadrierten euklidischen Distanzen**

(4.5) $$QED(i,j) = \sum_{k=1}^{m} (z_{ik} - z_{jk})^2$$

Gebrauch. Die Rangfolge der ähnlichen Objektpaare bleibt bei einer Verwendung der quadrierten euklidischen Distanz anstelle der euklidischen Distanz unverändert. Allerdings verändern sich die Abstandsverhältnisse, was durchaus einen Einfluss auf eine exploratorisch zu ermittelnde Gruppenzahl haben kann.

**Beispiel 4.5**: Eine Distanzmatrix mit den quadrierten euklidischen Distanzen kann analog zum vorherigen Fall mit SPSS aufgerufen werden. Hierzu ist im Feld „Maße“ der Eintrag „Quadrierte Euklidische Distanz“ zu wählen.

**Näherungsmatrix**

| | Quadrierte Euklidische Distanz | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1:A | 2:B | 3:C | 4:D | 5:E | 6:F | 7:G | 8:H | 9:I | 10:J | 11:K | 12:L |
| 1:A | .000 | 19.693 | 9.509 | 5.102 | 4.670 | 11.060 | 5.608 | 10.845 | 10.351 | 5.871 | 10.005 | 4.628 |
| 2:B | 19.693 | .000 | 45.926 | 8.334 | 15.168 | 13.765 | 10.382 | 28.883 | 10.166 | 13.250 | 22.166 | 8.690 |
| 3:C | 9.509 | 45.926 | .000 | 18.823 | 17.038 | 26.950 | 18.927 | 10.423 | 28.868 | 15.095 | 13.667 | 19.010 |
| 4:D | 5.102 | 8.334 | 18.823 | .000 | 7.736 | 11.975 | 2.543 | 9.027 | 9.130 | 2.528 | 6.017 | 3.448 |
| 5:E | 4.670 | 15.168 | 17.038 | 7.736 | .000 | 1.721 | 8.816 | 21.368 | 1.849 | 7.177 | 19.832 | 1.595 |
| 6:F | 11.060 | 13.765 | 26.950 | 11.975 | 1.721 | .000 | 11.754 | 29.137 | 1.041 | 9.914 | 27.263 | 3.775 |
| 7:G | 5.608 | 10.382 | 18.927 | 2.543 | 8.816 | 11.754 | .000 | 7.108 | 11.138 | 1.968 | 6.767 | 5.285 |
| 8:H | 10.845 | 28.883 | 10.423 | 9.027 | 21.368 | 29.137 | 7.108 | .000 | 29.265 | 6.394 | 1.669 | 17.439 |
| 9:I | 10.351 | 10.166 | 28.868 | 9.130 | 1.849 | 1.041 | 11.138 | 29.265 | .000 | 10.039 | 26.230 | 2.107 |
| 10:J | 5.871 | 13.250 | 15.095 | 2.528 | 7.177 | 9.914 | 1.968 | 6.394 | 10.039 | .000 | 5.516 | 5.333 |
| 11:K | 10.005 | 22.166 | 13.667 | 6.017 | 19.832 | 27.263 | 6.767 | 1.669 | 26.230 | 5.516 | .000 | 15.227 |
| 12:L | 4.628 | 8.690 | 19.010 | 3.448 | 1.595 | 3.775 | 5.285 | 17.439 | 2.107 | 5.333 | 15.227 | .000 |

Dies ist eine Unähnlichkeitsmatrix

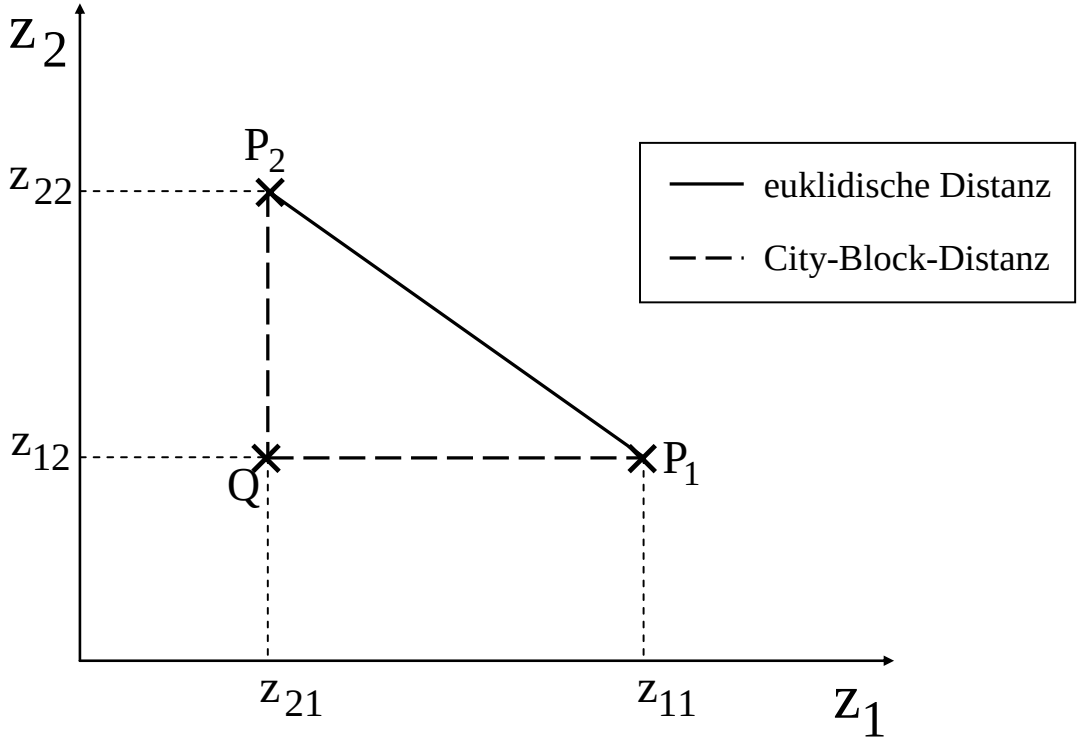
## City-Block-Distanz

Man gelangt in unserem angegebenen Koordinatensystem nicht allein auf dem "direkten Weg" vom Punkt $P_1$ zum Punkt $P_2$, sondern man kann auch einen "Umweg" über den Punkt Q nehmen. Von $P_1$ bis Q bewegt man sich hierbei parallel zur $x_1$-Achse, von Q bis P2 verläuft der Weg dann parallel zur $x_2$-Achse. Ein Distanzmaß, das auf den rechtwinkligen Abständen der Merkmalsvektoren in einem n-dimensionalen Zahlenraum basiert, heißt **City-Block-Distanz**:

(4.6) $$CBD(i,j) = \sum_{k=1}^{m} \left| z_{ik} - z_{jk} \right|$$

CBD: Summation der absoluten Merkmalswertdifferenzen

Abbildung 4.3: Euklidische Distanz und City-Block-Distanz in der Ebene



Bezeichnungen „**Manhattan- oder Taxi-Driver-Distanz“:**
Herstellung einer Verbindung zu den Fahrtwegen, die ein Taxifahrer zurückzulegen hat, um ein Ziel in einem rechtwinklig angeordneten Straßennetz eines Stadtviertels (Manhattan) zu erreichen

**Beispiel 4.6**: Unter Verwendung der beiden Merkmale Bevölkerungsdichte ($x_1$) und Wanderungssaldo ($x_6$) erhält man eine City-Block-Distanz zwischen den Regionen A und K in Höhe von

$$\begin{aligned} CBD(A,K) &= |z_{A1} - z_{K1}| + |z_{A6} - z_{K6}| \\ &= |(-0{,}657) - (-0{,}922)| + |(-0{,}470) - 0{,}931| \\ &= |0{,}266| + |-1{,}401| = 1{,}667. \end{aligned}$$

In der mit SPSS durch die Maß des Maßes „Block“ berechneten Distanzmatrix

**Näherungsmatrix**

| | City-Block-Distanzmaß | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1:A | 2:B | 3:C | 4:D | 5:E | 6:F | 7:G | 8:H | 9:I | 10:J | 11:K | 12:L |
| 1:A | .000 | 10.218 | 5.963 | 4.245 | 4.373 | 6.667 | 5.116 | 7.065 | 6.468 | 5.075 | 6.550 | 3.639 |
| 2:B | 10.218 | .000 | 16.181 | 6.196 | 8.618 | 6.923 | 7.363 | 11.956 | 5.743 | 8.120 | 9.503 | 6.578 |
| 3:C | 5.963 | 16.181 | .000 | 9.985 | 8.543 | 10.688 | 9.665 | 6.589 | 10.936 | 8.136 | 7.508 | 9.603 |
| 4:D | 4.245 | 6.196 | 9.985 | .000 | 4.958 | 6.991 | 3.649 | 6.673 | 6.942 | 2.529 | 5.467 | 3.989 |
| 5:E | 4.373 | 8.618 | 8.543 | 4.958 | .000 | 2.657 | 5.651 | 10.925 | 2.916 | 5.206 | 10.314 | 2.570 |
| 6:F | 6.667 | 6.923 | 10.688 | 6.991 | 2.657 | .000 | 7.248 | 13.070 | 2.115 | 7.253 | 12.459 | 4.597 |
| 7:G | 5.116 | 7.363 | 9.665 | 3.649 | 5.651 | 7.248 | .000 | 5.822 | 7.198 | 2.938 | 5.928 | 4.720 |
| 8:H | 7.065 | 11.956 | 6.589 | 6.673 | 10.925 | 13.070 | 5.822 | .000 | 13.020 | 5.817 | 2.632 | 9.738 |
| 9:I | 6.468 | 5.743 | 10.936 | 6.942 | 2.916 | 2.115 | 7.198 | 13.020 | .000 | 7.203 | 12.409 | 3.282 |
| 10:J | 5.075 | 8.120 | 8.136 | 2.529 | 5.206 | 7.253 | 2.938 | 5.817 | 7.203 | .000 | 5.206 | 4.213 |
| 11:K | 6.550 | 9.503 | 7.508 | 5.467 | 10.314 | 12.459 | 5.928 | 2.632 | 12.409 | 5.206 | .000 | 9.127 |
| 12:L | 3.639 | 6.578 | 9.603 | 3.989 | 2.570 | 4.597 | 4.720 | 9.738 | 3.282 | 4.213 | 9.127 | .000 |

Dies ist eine Unähnlichkeitsmatrix

sind die City-Block-Distanzen für alle zwölf Regionen aus sechs Merkmalen wiedergegeben. Die Rangfolge der ähnlichen Objektpaare verändert sich hierbei gegenüber der euklidischen Distanz. Zwar erweisen sich auf der Grundlage der City-Block-Distanz nach wie vor die Regionen F und I am ähnlichsten, aber den Rang 2 nehmen jetzt z.B. die Regionen D und J ein, während hier vorher die Regionen E und L platziert waren. Erneut sei der komplette Distanzwert für die Regionen A und K hier rechnerisch nachvollzogen:

$$
\begin{aligned}
\mathrm{CBD}(A,K) &= \sum_{k=1}^{6} \left| z_{Ak} - z_{Kk} \right| \\
&= \left|(-0{,}657)-(-0{,}922)\right| + \left|(-1{,}245)-(-0{,}940)\right| + \left|(-0{,}073)-(-0{,}713)\right| \\
&+ \left|(-0{,}450)-1{,}631\right| + \left|(-0{,}450)-1{,}262\right| + \left|(-0{,}470)-(-0{,}931)\right| \\
&= \left|0{,}226\right| + \left|-0{,}305\right| + \left|-0{,}786\right| + \left|-2{,}081\right| + \left|-1{,}712\right| + \left|-1{,}401\right| \\
&= 6{,}550
\end{aligned}
$$

♦

## 4.2.2 Ähnlichkeitskoeffizienten

### A. Dichotome Merkmale

**Dichotome Merkmale**:
Zweiwertige Merkmale, deren Modalitäten hier mit 0 und 1 kodiert werden (binäre Kodierung)
Sofern eine bestimmte Eigenschaft im Blickpunkt des Interesses steht, wird ihr der Wert 1 zugeordnet. Bei gleichwertigen Eigenschaften ist die Kodierung beliebig. Die Ähnlichkeitsmaße lassen sich dann auf der Basis einer Vierfeldertafel formal eindeutig und nachvollziehbar definieren.

Vierfeldertafel

| Objekt i | Objekt j | | $\Sigma$ |
|---|---|---|---|
| | 1 | 0 | |
| 1 | a | b | a + b |
| 0 | c | d | c + d |
| $\Sigma$ | a + c | b + d | a + b + c + d = m |

**Vierfeldertafel**:
Kombinationen von m dichotomen Merkmalen für die beiden Objekte i und j
Häufigkeit a:
Gibt an, wie oft die interessierenden Eigenschaften der Merkmale bei beiden Objekten gleichzeitig vorhanden sind

Häufigkeit d:
gibt an, wie oft diese Eigenschaften bei beiden Objekten zugleich nicht vorhanden sind
Häufigkeiten b und c:
erfassen Besetzungszahlen der Kombinationen, bei denen jeweils eine Modalität eines Merkmals bei einem Objekt vorhanden ist, nicht jedoch bei dem anderen Objekt
Bei m Merkmalen muss die Summe der Größen a, b, c und d stets gleich m sein:

$$a + b + c + d = m.$$

Auf der Basis der **Vierfeldertafel** lassen sich **unterschiedliche Ähnlichkeitsmaße** konzipieren.

## Simple-Matching-Koeffizient (SMC)

Der **Simple-Matching-Koeffizient (SMC)** setzt die gleichzeitig vorhandenen Eigenschaften bei den beiden Objekten i und j zur Gesamtzahl der zum Vergleich herangezogenen Merkmale in Beziehung:

(4.7) $$SMC(i,j) = \frac{a+d}{m}$$

Hierbei wird nicht zwischen dem Vorhandensein einer ganz bestimmten Eigenschaft wie z.B. Produktkauf oder Erwerbstätigkeit unterschieden. Diese Merkmalsausprägungen werden genauso behandelt wie ein Nicht-Kauf oder die Nicht-Erwerbstätigkeit.

## Jaccard-Koeffizient

Dagegen stellt der **Jaccard-Koeffizient** allein auf die potenziell interessierenden Eigenschaften ab. Die Merkmale, bei denen die im Blickpunkt stehende Eigenschaft bei beiden Objekten nicht vorhanden ist, werden bei der Ähnlichkeitsmessung von vornherein ignoriert. Der Jaccard-Koeffizient (J) ist dann durch

(4.8) $$J(i,j) = \frac{a}{a+b+c}$$

definiert. In diesem Fall wird eine Ähnlichkeit zwischen den beiden Objekten i und j allein z.B. aus dem Produktkauf oder der Erwerbstätigkeit, nicht jedoch aus dem Nicht-Kauf oder der Nicht-Erwerbstätigkeit abgeleitet.

## RR-Koeffizient

Beim **RR-Koeffizienten** vom Russel und Rao wird zwar ebenfalls eine Assoziation der Ähnlichkeit von Objekten mit bestimmten interessierenden Eigenschaften gegeben, jedoch wird im Unterschied zum Jaccard-Koeffizienten der Bezug zu allen betrachteten Merkmalen beibehalten:

(4.9) $$RR(i,j) = \frac{a}{m}$$

<u>Eigenschaft:</u>
Alle drei Ähnlichkeitsmaße liegen zwischen 0 (=minimale Ähnlichkeit) und 1 (=maximale Ähnlichkeit)

**Beispiel 4.7:** Bei acht Konsumenten sind vier dichotome Merkmale

- Produktkauf (ja, nein),
- Erwerbstätigkeit (ja, nein),
- soziale Stellung (hoch, niedrig),
- Bildung (hoch, niedrig)

erhoben worden. Kodiert man die potenziell interessierende Eigenschaft (ja bzw. hoch) mit 1 und das Nicht-Vorhandensein dieser Eigenschaft (nein bzw. niedrig) mit 0, dann lassen sich die Erhebungsdaten in Form der Beobachtungsmatrix

$$
\begin{array}{cc}
 & \begin{array}{cccc} \text{Produktkauf} & \text{Erwerbstätigkeit} & \text{Soziale Stellung} & \text{Bildung} \\ (X1) & (X2) & (X3) & (X4) \end{array} \\
\mathbf{X} = & \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{array}{l} \text{Konsumt 1} \\ \text{Konsumt 2} \\ \text{Konsumt 3} \\ \text{Konsumt 4} \\ \text{Konsumt 5} \\ \text{Konsumt 6} \\ \text{Konsumt 7} \\ \text{Konsumt 8} \end{array}
\end{array}
$$

wiedergeben.

Für die beiden Konsumenten 1 und 2 lässt sich hieraus eine Vierfeldertafel

| Konsument 1 | Konsument 2 | | |
|---|---|---|---|
| | 1 | 0 | Σ |
| 1 | a = 1 | b = 2 | a + b = 3 |
| 0 | c = 0 | d = 1 | c + d = 1 |
| Σ | a + c = 1 | b + d = 3 | m = 4 |

erstellen. Auf der Grundlage der Vierfeldertafel ergeben sich für die Ähnlichkeitsmaße (4.7) - (4.9) die Werte

$$SMC(1,2) = \frac{a+d}{m} = \frac{1+1}{4} = 0{,}5\,,$$

$$J(1,2) = \frac{a}{a+b+c} = \frac{1}{1+2+0} = 0{,}333 \quad \text{und}$$

$$RR(1,2) = \frac{a}{m} = \frac{1}{4} = 0{,}25\,.$$

Die unterschiedlichen Werte der Ähnlichkeitsmaße erklären sich aus verschiedenartigen Behandlung der Eigenschaft mit der Modalität 0. Während der Simple-Matching-Koeffizient mit 0,5 den größten Wert annimmt, ist der RR-Koeffizient von 0,25 am niedrigsten. Dazwischen liegt erwartungsgemäß der Jaccard-Koeffizient.

Die **Simple-Matching-Koeffizienten** aller acht Konsumenten sind wiederum mit SPSS berechnet worden. Hierzu wählen wir

Analysieren
    Korrelation
        Distanzen …

In dem Fenster „Distanzen“ klicken wir in dem Feld „Maße“ das Item „Ähnlichkeiten“ an. Anschließend klicken wir auf die Schaltfläche „Maße“. In dem sich öffnenden Fenster „Distanzen: Ähnlichkeitsmaße“ wählen wir im Feld „Maße“ das Item „Binär:“ sowie aus dem Pull-down-Menü den Eintrag „Einfache Übereinstimmung“.
Die **Ähnlichkeitsmatrix** (SPSS-Tabelle Näherungsmatrix) ist wiederum symmetrisch und enthält auf der Hauptdiagonale Einsen, die die perfekte Ähnlichkeit eines Vergleichs der Konsumenten mit sich selbst widerspiegeln.

**Näherungsmatrix**

| | Ähnlichkeitsmaß der einfachen Übereinstimmung | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 1.000 | .500 | .250 | .750 | .500 | .500 | .750 | .250 |
| 2 | .500 | 1.000 | .250 | .750 | 1.000 | .500 | .250 | .750 |
| 3 | .250 | .250 | 1.000 | .000 | .250 | .750 | .500 | .500 |
| 4 | .750 | .750 | .000 | 1.000 | .750 | .250 | .500 | .500 |
| 5 | .500 | 1.000 | .250 | .750 | 1.000 | .500 | .250 | .750 |
| 6 | .500 | .500 | .750 | .250 | .500 | 1.000 | .750 | .250 |
| 7 | .750 | .250 | .500 | .500 | .250 | .750 | 1.000 | .000 |
| 8 | .250 | .750 | .500 | .500 | .750 | .250 | .000 | 1.000 |

Dies ist eine Ähnlichkeitsmatrix

Um die **Ähnlichkeitsmatrix** auf der Basis des **Jaccard-Koeefizienten** zu erhalten wählen wir im Pull-down-Menü des SPSS-Fensters „Distanzmaße: Ähnlichkeits-maße“ den Eintrag „Jaccard“:

**Näherungsmatrix**

| | Ähnlichkeitsmaß nach Jaccard | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 1.000 | .333 | .250 | .667 | .333 | .500 | .750 | .000 |
| 2 | .333 | 1.000 | .000 | .500 | 1.000 | .333 | .250 | .000 |
| 3 | .250 | .000 | 1.000 | .000 | .000 | .667 | .500 | .000 |
| 4 | .667 | .500 | .000 | 1.000 | .500 | .250 | .500 | .000 |
| 5 | .333 | 1.000 | .000 | .500 | 1.000 | .333 | .250 | .000 |
| 6 | .500 | .333 | .667 | .250 | .333 | 1.000 | .750 | .000 |
| 7 | .750 | .250 | .500 | .500 | .250 | .750 | 1.000 | .000 |
| 8 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | 1.000 |

Dies ist eine Ähnlichkeitsmatrix

Schließlich wählen wir im Pull-down-Menü des SPSS-Fensters „Distanzen: Ähnlichkeitssmaße“ den Eintrag „Russel und Rao“, um die **Ähnlichkeitsmatrix** des **RR-Koeffizienten** zu erzeugen:

**Näherungsmatrix**

| | Ähnlichkeitsmaß nach Russell und Rao | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | .750 | .250 | .250 | .500 | .250 | .500 | .750 | .000 |
| 2 | .250 | .250 | .000 | .250 | .250 | .250 | .250 | .000 |
| 3 | .250 | .000 | .500 | .000 | .000 | .500 | .500 | .000 |
| 4 | .500 | .250 | .000 | .500 | .250 | .250 | .500 | .000 |
| 5 | .250 | .250 | .000 | .250 | .250 | .250 | .250 | .000 |
| 6 | .500 | .250 | .500 | .250 | .250 | .750 | .750 | .000 |
| 7 | .750 | .250 | .500 | .500 | .250 | .750 | 1.000 | .000 |
| 8 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |

Dies ist eine Ähnlichkeitsmatrix

Allgemein kann das V**erhältnis der Ähnlichkeitsmaße** (4.7) – (4.9) zueinander durch die Ungleichung $RR(i,j) \leq J(i,j) \leq SMC(i,j)$

wiedergegeben werden. Je nach Wahl des Ähnlichkeitsmaßes durch den Anwender kann sich die Ähnlichkeit zwischen Objekten graduell unterschiedlich darstellen. Die drei Maße sind nur dann identisch, wenn die im Blickpunkt stehenden Eigenschaften nicht gleichzeitig bei beiden Objekten vorhanden sind. In der Anwendung ist daher stets zu beurteilen, ob die Ähnlichkeit zwischen den Objekten sachgerecht immer das Vorhandensein einer bestimmten Eigenschaft voraussetzt oder ob die Modalitäten der Untersuchungsmerkmale gleichwertig zu behandeln sind. ♦

## B. Polytome Merkmale

**Polytome Merkmale:**

Mehrkategorielle Merkmale, die auf einem nominalen oder ordinalen Skalenniveau gemessen werden können

Grundsätzlich kann eine **Ähnlichkeitsmessung bei polytomen Merkmalen** auf der Basis der für dichotome Merkmale entwickelten Ähnlichkeitskonzepte nach Vornahme einer **Dichotomisierung** erfolgen.

Merkmale mit zwei Kategorien:

1 Dummy-Variable mit den Werten 0 und 1

Merkmale mit 3 Kategorien:

Zwei 0,1-Dummy-Variablen

**Beispiel 4.8:** Wir betrachten das **Merkmal "Soziale Stellung"** ($X_3$), das bisher zweiwertig definiert worden ist:

$$x_{i3} = \begin{cases} 1, \text{ falls "hoch"} \\ 0 \text{ sonst}. \end{cases}$$

Da nur zwischen einer **hohen und niedrigen sozialen Stellung** unterschieden worden ist, gibt der Wert 0 eine niedrige soziale Stellung einer Untersuchungseinheit (=Konsument) i wieder.

Wenn wir jetzt zusätzlich die **Kategorie "mittel"** der sozialen Stellung betrachten, lässt sich das Merkmal "Soziale Stellung" in eine 0,1-Kodierung durch zwei Dummy-Variablen erfassen:

$$x'_{i3} = \begin{cases} 1, \text{ falls "mittel"} \\ 0 \text{ sonst} \end{cases}$$

und

$$x'_{i4} = \begin{cases} 1, \text{ falls "hoch"} \\ 0 \text{ sonst}. \end{cases}$$

**Referenzkategorie:** ***Ausprägung "niedrig"***, [ Anmerkung: Eine der beiden anderen Kategorien käme hier genauso gut in Betracht. ] Aus dem **polytomen Merkmal "Soziale Stellung"** ($X_3$) sind quasi **zwei neue dichotome Merkmale "Mittlere soziale Stellung"** ($X'_3$) und **"Hohe soziale Stellung"** ($X'_4$) entstanden.

Konsument mit einer niedrigen sozialen Stellung:
Ausprägungen $x'_3 = 0$ und $x'_4 = 0$ bei den beiden neuen Merkmalen $X'_3$ und $X'_4$ ♦

Wie unmittelbar einsichtig ist, kann allgemein ein **Merkmal** mit **p Kategorien** durch **p-1 Dummyvariablen** wiedergegeben werden.

**Gewichtungsproblem:**
Eine unmodifizierte Verwendung der für dichotome Merkmale konzipierten Ähnlichkeitsmaße ist keinesfalls unproblematisch. Das Gewicht eines originären Merkmals nimmt nämlich offenbar mit der Anzahl der Kategorien bei der Ähnlichkeitsmessung zu. Der Stellenwert eines Merkmals würde dann allein durch seine Kategorienzahl bestimmt werden. Hinzu kommt, dass "künstliche" Übereinstimmungen zwischen den zu klassifizierenden Objekten erzeugt werden, die bestimmte Ähnlichkeitskoeffizienten nach oben verzerren.

**Beispiel 4.9: 3-kategorielles Merkmal "Soziale Stellung"**
Angenommen, die soziale Stellung einer Untersuchungseinheit ist niedrig und diejenige einer anderen Untersuchungseinheit ist hoch. Dann würde sich für diese beiden Untersuchungseinheiten **eine Übereinstimmung** in dem neuem Merkmal "Mittlere soziale Stellung" ($X'_3$) ergeben, da sie hierin einheitlich die Ausprägung 0 erhielten. Würde die soziale Stellung z.B. auf **fünf Kategorien** verfeinert, erhielte man für die Untersuchungseinheiten, die unterschiedlichen Kategorien angehörten, bereits **drei Übereinstimmungen**. ♦

Um den Stellenwert der originären Merkmale zu erfassen, ist eine Gewichtung der Dummy-Variablen polytomer Merkmale vorzunehmen. Da ein **polytomes Merkmal** mit p Kategorien durch **p-1 Dummy-Variablen** abgebildet wird, erhält jede **Dummy-Variable** das **Gewicht** 1/(p-1).

**Beispiel 4.10:** Neben der **sozialen Stellung** sei ebenfalls das **Merkmal Bildung** durch **drei Kategorien niedrig, mittel, hoch** erfasst. Für die acht Konsumenten liegt dann die Beobachtungsmatrix in der Form

| Gewicht: | 1 | 1 | ½ | ½ | ½ | ½ | |
|---|---|---|---|---|---|---|---|
| | Produkt-kauf ($X_1$) | Erwerbs-tätigkeit ($X_2$) | Mittl. soz. Stellung ($X_3$) | Hohe soz. Stellung ($X_4$) | Mittlere Bildung ($X_5$) | Hohe Bildung ($X_6$) | |
| $\mathbf{X} =$ | 1 | 1 | 0 | 1 | 1 | 0 | Konsument 1 |
| | 1 | 0 | 0 | 0 | 1 | 0 | Konsument 2 |
| | 0 | 1 | 1 | 0 | 0 | 1 | Konsument 3 |
| | 1 | 0 | 0 | 1 | 1 | 0 | Konsument 4 |
| | 1 | 0 | 1 | 0 | 0 | 0 | Konsument 5 |
| | 1 | 1 | 1 | 0 | 0 | 1 | Konsument 6 |
| | 1 | 1 | 0 | 1 | 0 | 1 | Konsument 7 |
| | 0 | 0 | 0 | 0 | 1 | 0 | Konsument 8 |

vor.

Bei der Berechnung der Größen a, b, c und d der Vierfeldertafel sind die von vornherein dichotomen Merkmale Produktkauf und Erwerbstätigkeit weiterhin mit 1 zu gewichten. Dagegen erhalten die aus den beiden polytomen Merkmalen (k = 3) Soziale Stellung und Bildung hervorgegangenen Dummy-Variablen "Mittlere und hohe soziale Stellung" sowie "Mittlere und hohe Bildung" jeweils das Gewicht ½.

Für die Konsumenten 1 und 2 nehmen die Größen a, b, c und d daher die Werte

$$a = 1 + \frac{1}{2} \cdot 1 = 1{,}5, \quad b = 1 + \frac{1}{2} \cdot 1 = 1{,}5, \quad c = 0 \text{ und } \quad d = \frac{1}{2} \cdot 2 = 1$$

an. Hieraus erhält man die Ähnlichkeitskoeffizienten

$$SMC(1,2) = \frac{a+d}{a+b+c+d} = \frac{1{,}5+1}{4} = 0{,}625 \ ,$$

$$J(1,2) = \frac{a}{a+b+c} = \frac{1{,}5}{1{,}5+1{,}5+0} = 0{,}500 \quad \text{und}$$

$$RR(1,2) = \frac{a}{a+b+c+d} = \frac{1{,}5}{4} = 0{,}375 \ .$$

♦