G. Rohwer, C. Dudel Wintersemester 2010/11

Methoden der Datenrepräsentation und Klassifikation

Kapitel 2: Abstandskonstruktionen

2 Abstandskonstruktionen

Viele der in diesem Text besprochenen Methoden beziehen sich auf Abstände zur Erfassung von Ähnlichkeiten oder anderen Arten von Beziehungen. In einigen Fällen können Abstände unmittelbar erhoben werden; beispielsweise in Form von Daten über soziale Netzwerke oder wenn unmittelbar subjektive Ähnlichkeitsurteile erfragt werden.[1] Andererseits können Abstandsfunktionen auch aus anderen Arten von Daten, insbesondere aus statistischen Variablen, konstruiert werden. Einige Möglichkeiten werden in diesen Kapitel besprochen. Die Überlegungen erfolgen anhand von Beispielen, die in späteren Kapiteln zur Illustration von Verfahren der Datenrepräsentation und Klassifikation dienen.

[1] Zur direkten Erfragung von Ähnlichkeitsurteilen vgl. man beispielsweise Green, Carmone und Smith (1989: 60ff.).

2.1 Abstände zwischen Merkmalswerten

1. Objekte und Merkmalswerte

Wir unterscheiden zwischen Objekten und Merkmalswerten. Zwar könnte der Objektbegriff so weit gefasst werden, dass man auch Merkmalswerte als abstrakte Objekte auffassen kann; dennoch bleibt eine Unterscheidung: Konkrete oder abstrakte Objekte gehören zum jeweils thematisierten Gegenstandsbereich; Merkmalswerte dienen dazu, die konkreten oder abstrakten Objekte zu charakterisieren.[2] Wichtig ist die Unterscheidung auch deshalb, weil Abstandskonstruktionen entweder bei Objekten oder bei Merkmalswerten ansetzen können.

Ein Beispiel wurde bereits in Abschnitt 1.2 (§ 3) angeführt: Schulabschlüsse. In diesem Beispiel kann man einerseits Abstände zwischen Schulabschlüssen definieren; dann konstruiert man eine Abstandsfunktion für einen Merkmalsraum (dessen Elemente begrifflich definierte Schulabschlüsse sind). Andererseits kann man auch unter Bezugnahme auf Schulabschlüsse Abstände zwischen Menschen definieren; dann konstruiert man eine Abstandsfunktion für Objekte (Menschen).

2. Definitionen für Merkmalsräume

Bei der Bildung von Abstandsfunktionen für Merkmalsräume kann man auf zwei wesentlich unterschiedliche Weisen vorgehen.

- Man kann die Bildung einer Abstandsfunktion für einen Merkmalsraum als eine theoretische Aufgabe ansehen, die grundsätzlich unabhängig von Daten erfolgen sollte (was natürlich empirische Bezüge nicht ausschließt). Abstandsdefinitionen sollten dann insbesondere nicht von den Häufigkeiten abhängen, mit denen die Merkmalswerte in irgendwelchen empirisch fixierbaren Gesamtheiten vorkommen. Folgt man dieser Vorgehensweise, sollten zum Beispiel Abstände zwischen Schulabschlüssen unabhängig davon definiert werden, mit welchen Häufigkeiten die unterschiedlichen Schulabschlüsse vorkommen.
- Andererseits gibt es daten- bzw. verteilungsabhängige Verfahren der Abstandskonstruktion für Merkmalsräume. Die resultierenden Abstandsfunktionen hängen dann auf kontingente Weise von den Daten ab, die für die Konstruktion verwendet werden.

[2] Wir versuchen in diesem Text, auch sprachlich zwischen Objekten und Merkmalswerten zu unterscheiden. Mit dem Objektbegriff beziehen wir uns meistens auf eine der folgenden Arten:

a) *Individuelle Objekte*, die empirisch identifiziert werden können; zum Beispiel: Menschen, Tiere, Häuser.

b) *Institutionelle Objekte*, zum Beispiel: Haushalte, Unternehmen, Länder.

c) *Statistisch konstruierte Objekte*, zum Beispiel: Berufe, statistische Verteilungen.

In der Kategorie (c) gibt es scheinbar eine Überschneidung mit Merkalswerten, denn beispielsweise Berufe können auch als Merkmalswerte verwendet werden. Werden Berufe jedoch als statistisch konstruierte Objekte verwendet, sind jeweils Gesamtheiten von Menschen gemeint, die einen Beruf innehaben bzw. ausüben.

Tabelle 2.1-1 Durch die Kemeny-Metrik definierte Abstände für Rangordnungen mit drei Alternativen.

		1	2	3	4	5	6	7	8	9	10	11	12	13
1	(1,2,3)	0	1	2	3	4	5	6	5	4	3	2	1	3
2	(1,1,3)	1	0	1	2	3	4	5	6	5	4	3	2	2
3	(2,1,3)	2	1	0	1	2	3	4	5	6	5	4	3	3
4	(2,1,2)	3	2	1	0	1	2	3	4	5	6	5	4	2
5	(3,1,2)	4	3	2	1	0	1	2	3	4	5	6	5	3
6	(3,2,2)	5	4	3	2	1	0	1	2	3	4	5	6	2
7	(3,2,1)	6	5	4	3	2	1	0	1	2	3	4	5	3
8	(2,2,1)	5	6	5	4	3	2	1	0	1	2	3	4	2
9	(2,3.1)	4	5	6	5	4	3	2	1	0	1	2	3	3
10	(2,3,2)	3	4	5	6	5	4	3	2	1	0	1	2	2
11	(1,3,2)	2	3	4	5	6	5	4	3	2	1	0	1	3
12	(1,2,2)	1	2	3	4	5	6	5	4	3	2	1	0	2
13	(2,2,2)	3	2	3	2	3	2	3	2	3	2	3	2	0

Wir sprechen im ersten Fall von *verteilungsunabhängigen*, im zweiten Fall von *verteilungsabhängigen* Verfahren der Abstandskonstruktion.[3]

Zur Verdeutlichung einer verteilungsunabhängigen Abstandskonstruktion beziehen wir uns auf einen Merkmalsraum für Rangordnungen mit drei Alternativen. Es handelt sich um einen qualitativen Merkmalsraum. Gleichwohl ist es möglich, eine Abstandsfunktion zu definieren. Dafür kann die sog. *Kemeny-Metrik* verwendet werden (Rohwer und Pötter 2002a: Kap. 11). Zum Beispiel gibt es bei drei Alternativen insgesamt 13 Rangordnungen, zwischen denen mit der Kemeny-Metrik Abstände berechnet werden können. Abbildung 2.1-1 zeigt für dieses Beispiel die Abstandsmatrix.

Offenbar handelt es sich um eine verteilungsunabhängige Abstandsdefinition, denn es wird nur auf Eigenschaften der Rangordnungen Bezug genommen, vollständig unabhängig davon, wie häufig die Rangordnungen in irgendeinem Anwendungsfall auftreten.

2.2 Abstände zwischen Objekten

1. Verwendung von Merkmalswerten

Abstände zwischen Objekten können entweder direkt empirisch ermittelt oder aus Merkmalswerten der Objekte konstruiert werden. Hier beschäftigen wir uns mit der zweiten Möglichkeit. Gibt es bereits eine Abstandsfunktion für den Merkmalsraum, kann man für die Objekte eine indu-

[3] Zu einer analogen Unterscheidung vgl. man Rohwer und Pötter (2002a: 71f.).

zierte Abstandsfunktion verwenden (vgl. Abschnitt 1.2, § 4). Oft besteht die Aufgabe jedoch darin, zur Bildung von Abständen gleichzeitig mehrere Merkmalswerte heranzuziehen; und selbst wenn es für jeden einzelnen Merkmalsraum eine Abstandsfunktion gibt, ergibt sich daraus nicht ohne weiteres eine Abstandsfunktion für den gemeinsamen Merkmalsraum.

Hier setzen Vorschläge an, wie man unter Verwendung mehrerer Merkmalswerte Abstandsfunktionen definieren kann. Es gibt sehr viele solcher Vorschläge, deren Anwendungsmöglichkeiten auch davon abhängen, von welcher Art die beteiligten Merkmalsräume sind.[4]

2. Notationen für Datenmatrizen

Um einige der Standarddefinitionen zu erläutern, beziehen wir uns auf eine statistische Variable mit m Komponenten:

$$(X_1, \ldots, X_m) : \Omega \longrightarrow \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$$

Wenn die Elemente von Ω durch $i = 1, \ldots, n$ numeriert werden, können die Daten durch folgende Datenmatrix erfasst werden:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nm} \end{pmatrix}$$

Jede Zeile entspricht einem Objekt; x_{ik} ist der Merkmalswert des Objekts ω_i bei der Variablen X_k bzw. im Merkmalsraum $\mathcal{X}_k$.[5]

Die Aufgabe besteht darin, Abstände für die Zeilen der Matrix $\mathbf{X}$ zu definieren.[6] Jede Zeile kann als ein Vektor aufgefasst werden, der aus m Zahlen besteht. Da wir Vektoren stets als Spaltenvektoren betrachten, verwenden wir folgende Notation:

$$\mathbf{x}_i := (x_{i1}, \ldots, x_{im})' \qquad (i = 1, \ldots, n)$$

Um eine Abstandsfunktion zu definieren, muss eine Funktion angegeben werden, die für alle Paare von Zeilen $\mathbf{x}_i$ und $\mathbf{x}_j$ eine Zahl $d(\mathbf{x}_i, \mathbf{x}_j)$ angibt, die als ein Abstand interpretiert werden kann und den in Abschnitt 1.2 (§ 2) angegebenen Bedingungen genügt.

3. Abstände für Datenmatrizen

Aus der großen Anzahl der in der Literatur vorgeschlagenen Definitionen von Abstandsfunktionen für statistische Datenmatrizen beschränken wir

[4] Übersichten findet man beispielsweise bei Bock (1974: Teil I); Fox (1982); Cox und Cox (1994: 8ff.); Bacher (1994: 198ff.); Batagelj und Bren (1995).

[5] Es sei an die Annahme erinnert, dass es für Merkmalswerte stets eine numerische Repräsentation gibt. Bei den Koeffizienten der Datenmatrix $\mathbf{X}$ handelt es sich also um Zahlen.

[6] Eine formal analoge, aber inhaltlich andere Fragestellung betrifft Zusammenhänge zwischen den Spalten einer Datenmatrix. Ein Beispiel wird in § 6 besprochen.

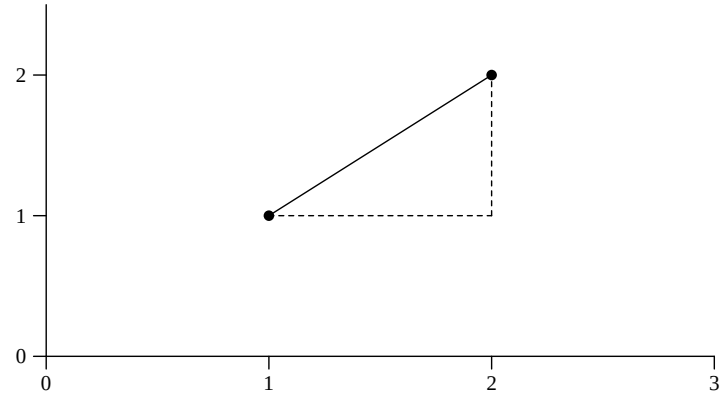

Abb. 2.2-1 Vergleich des euklidischen und des City-Block-Abstands zwischen den Punkten $(1, 1)$ und $(2, 2)$.

uns hier auf einige der am häufigsten verwendeten.[7] Wir verwenden die eben erläuterte Notation und beziehen uns auf zwei Zeilen $\mathbf{x}_i$ und $\mathbf{x}_j$ der Datenmatrix $\mathbf{X}$.

- Der *euklidische Abstand*

  $$\| \mathbf{x}_i - \mathbf{x}_j \| := \left(\sum\nolimits_{k=1,m} (x_{ik} - x_{jk})^2 \right)^{1/2}$$

  Da diese Definition die Bedingungen für eine Metrik erfüllt, wird auch von einer *euklidischen Metrik* gesprochen.

- Die *Summe der absoluten Differenzen*

  $$d^a(\mathbf{x}_i, \mathbf{x}_j) := \sum\nolimits_{k=1,m} |x_{ik} - x_{jk}|$$

  Auch in diesem Fall handelt es sich um eine Metrik; sie wird auch *City-Block-Metrik* genannt.

- Wenn alle beteiligten Variablen nur zwei Merkmalswerte (0 und 1) aufweisen können, wird der City-Block-Abstand auch als *Hamming-Distanz* bezeichnet. Der Abstand erfasst dann einfach die Anzahl der Merkmale, in denen $\mathbf{x}_i$ und $\mathbf{x}_j$ voneinander abweichen.

Eine graphische Darstellung des euklidischen und des City-Block-Abstands zwischen zwei Punkten $(1, 1)$ und $(2, 2)$ findet man in Abbildung 2.2-1. Der euklidische Abstand beträgt in diesem Beispiel $\sqrt{2}$ und entspricht der

[7] Eine Auswahl weiterer Abstandsfunktionen findet man beispielsweise bei Bortz (2005: 567ff.).

Länge einer direkten Verbindung der Punkte. Der City-Block-Abstand hat dagegen den Wert 2 (Länge der gestrichelten Linie). Für zwei Punkte $(1, 1)$ und $(1, 2)$ wären die beiden Abstände gleich.

4. Gruppierte Daten und Abstände

Die Abstände zwischen identischen Zeilen einer Datenmatrix sind offenbar Null, und ihre Abstände zu allen übrigen Zeilen sind identisch. Es ist deshalb sinnvoll, eine Datenmatrix zunächst in eine gruppierte Form zu bringen, bevor aus ihr eine Abstandsmatrix erzeugt wird. Damit ist gemeint, dass eine neue Matrix gebildet wird, die nur unterschiedliche Zeilen enthält und außerdem eine Angabe über die Häufigkeit, mit der jede Zeile in der ursprünglichen Datenmatrix vorkommt. Man spricht dann auch von *gruppierten Daten*. Dieses Vorgehen ist insbesondere dann sinnvoll, wenn Abstände zwischen einer großen Anzahl an Objekten berechnet werden sollen, da hierdurch der Rechenaufwand ggf. erheblich reduziert werden kann.

5. Illustration mit Klausurdaten

Als ein erstes Beispiel verwenden wir die Ergebnisse einer Statistikklausur, an der 46 Personen teilgenommen haben.[8] Box 2.2-1 zeigt die Daten. Die erste Spalte enthält eine fortlaufende Nummer; dann folgen die Punktzahlen für fünf Aufgaben, in denen maximal 10, 10, 5, 10 und 15 Punkte erzielt werden konnten. Wie man sieht, treten einige Zeilen mehrfach auf; insgesamt gibt es 31 unterschiedliche Zeilen.

Wie kann man Abstände zwischen den Zeilen definieren? In diesem Beispiel liegt es nahe, einen additiven Index zu verwenden, also in jeder Zeile die Summe der Punkte zu berechnen und dann deren Differenzen als Abstände zu nehmen. Dann wäre es auch einfach, das Ergebnis darzustellen; man könnte einfach die Verteilung der Indexwerte durch eine Häufigkeitstabelle oder eine Verteilungsfunktion darstellen.

Allerdings ist dies nur eine von vielen möglichen Abstandskonstruktionen. Wenn man die Unterschiede in der Bearbeitung der fünf Klausuraufgaben erfassen möchte, könnte der City-Block-Abstand verwendet werden. Bereits beim Vergleich der ersten beiden Zeilen zeigt sich dann ein Unterschied. Der additive Index liefert in beiden Fällen 35 Punkte, so dass der Abstand Null ist. Verwendet man dagegen die City-Block-Metrik beträgt der Abstand 10.

Als ein Beispiel, das gelegentlich in späteren Kapiteln verwendet wird, erzeugen wir eine Abstandsmatrix mit den City-Block-Abständen. Es handelt sich um eine Matrix mit 31 Zeilen und Spalten; wir nennen sie die *City-Block-Abstandsmatrix für die Klausurdaten*. Box 2.2-2 zeigt ein R-Skript, das für die Berechnungen verwendet werden kann. Zu beachten ist,

[8] Diese Daten wurden bereits zur Illustration einiger Methoden der Datenkonstruktion bei Rohwer und Pötter (2002a: 76) verwendet.

Box 2.2-1 Klausurdaten (Datenfile `klaus1.dat`).

1	2	10	0	8	15	24	10	10	2	6	10
2	2	10	5	3	15	25	10	10	2	6	10
3	4	10	0	5	0	26	10	10	3	6	7
4	8	5	0	4	1	27	10	10	3	6	10
5	8	7	0	10	7	28	10	10	5	4	9
6	8	10	0	9	15	29	10	10	5	4	12
7	8	10	5	6	0	30	10	10	5	8	13
8	10	0	5	1	15	31	10	10	5	8	15
9	10	5	5	0	2	32	10	10	5	8	15
10	10	5	5	5	15	33	10	10	5	8	15
11	10	5	5	5	15	34	10	10	5	9	9
12	10	10	0	0	13	35	10	10	5	9	9
13	10	10	0	0	14	36	10	10	5	10	0
14	10	10	0	4	12	37	10	10	5	10	13
15	10	10	0	4	12	38	10	10	5	10	14
16	10	10	0	6	14	39	10	10	5	10	15
17	10	10	0	8	15	40	10	10	5	10	15
18	10	10	0	9	15	41	10	10	5	10	15
19	10	10	0	10	7	42	10	10	5	10	15
20	10	10	0	10	15	43	10	10	5	10	15
21	10	10	0	10	15	44	10	10	5	10	15
22	10	10	1	10	9	45	10	10	5	10	15
23	10	10	1	10	9	46	10	10	5	10	15

dass es in dieser Abstandsmatrix viele *Bindungen* gibt, d.h. viele Abstände kommen mehrfach vor. Tabelle 2.2-1 zeigt eine Häufigkeitsverteilung der Abstände.

6. Abstände zwischen Variablen

Statt Abstände zwischen den Zeilen kann man auch Abstände zwischen den Spalten (Variablen) einer Datenmatrix betrachten. In unserem Beispiel interessiert man sich dann für Abstände zwischen den Klausuraufgaben. Wiederum können solche Abstände auf viele unterschiedliche Weisen definiert werden.

Oft werden Korrelationskoeffizienten verwendet; Tabelle 2.2-2 zeigt sie für unser Beispiel (Box 2.2-3 zeigt den für die Berechnung verwendeten R-Befehl). Man erkennt zum Beispiel, dass eine relativ hohe Korrelation zwischen den Aufgaben 2 und 4 besteht. Korrelationskoeffizienten sind jedoch inhaltlich schwer zu interpretieren und liefern auch nicht unmittelbar Abstände. Ein großer Vorteil des Abstandsbegriffs liegt gerade darin, dass man Abstandsdefinitionen unter inhaltlichen Gesichtspunkten vornehmen kann; das wurde ja bereits im vorangegangenen Paragraphen deutlich. Dies trifft auch zu, wenn man sich für Unterschiede in der Art der Bearbeitung der Aufgaben interessiert. Man kann zum Beispiel aus den Klausurdaten in Box 2.2-1 folgende Häufigkeitstabelle bilden, die für

Box 2.2-2 R-Skript: Abstandsberechnungen mit Klausurdaten.

```
# Mit einer Raute (#) versehener Text ist ein Kommentar

# Annahme: Klausurdaten gespeichert unter C:\daten\klaus1.dat

setwd("C:/daten")
# Verzeichnis mit Daten waehlen
# Beachte: Anfuehrungszeichen notwendig; Slash anstatt Backslash

dat <- read.table("klaus1.dat")[,2:6]
# Klausurdaten einlesen
# nur Spaten 2 bis 6 der Daten eingelesen
# Daten als Objekt dat im Speicher und abrufbar
# Auch hier Anfuehrungszeichen notwendig

dat
# Datensatz aufrufen (Box 2.2-1)

dat_g <- unique(dat)
# Doppelte Zeilen ausschliessen

d <- dist(dat_g,method="manhattan")
# City-Block Metrik berechnen
# erstes Argument der Funktion dist sind Daten
# zweites Argument gibt die Abstandsfunktion an
# "manhattan" entspricht City-Block Metrik
# "euclidean" (Standard, muss nicht extra angegeben werden)
# entspricht euklidischer Metrik
# weitere Optionen mit help(dist) anzeigbar

table(d)
# Tabelle 2.2-1
```

Tabelle 2.2-1 Häufigkeiten der Abstände im unteren Dreieck der aus den gruppierten Klausurdaten gebildeten City-Block-Abstandsmatrix.

Abstand	Häufigkeit	Abstand	Häufigkeit	Abstand	Häufigkeit
1	6	13	27	25	10
2	6	14	16	26	7
3	8	15	25	27	4
4	5	16	22	28	6
5	12	17	22	29	8
6	15	18	19	30	6
7	21	19	18	31	6
8	21	20	14	32	4
9	19	21	19	33	1
10	33	22	10	34	2
11	18	23	13	39	1
12	26	24	14	40	1

Box 2.2-3 R-Skript: Berechnung von Korrelationskoeffizienten (Tabelle 2.2-2).

```
# im Anschluss an das Skript in Box 2.2-2:
cor(dat)
```

Tabelle 2.2-2 Korrelationen zwischen den Spalten (Aufgaben) des Klausurdatenfiles in Box 2.2-1.

Aufgabe 1	1.0000	-0.0320	0.1985	0.1458	0.1521
Aufgabe 2	-0.0320	1.0000	-0.1254	0.4434	0.1055
Aufgabe 3	0.1985	-0.1254	1.0000	0.1229	0.1175
Aufgabe 4	0.1458	0.4434	0.1229	1.0000	0.1962
Aufgabe 5	0.1521	0.1055	0.1175	0.1962	1.0000

jede Aufgabe angibt, wie gut sie gelöst worden ist:

$$\begin{pmatrix} 39 & 4 & 0 & 1 & 2 \\ 40 & 1 & 4 & 0 & 1 \\ 25 & 0 & 2 & 2 & 17 \\ 21 & 6 & 9 & 6 & 4 \\ 27 & 6 & 8 & 0 & 5 \end{pmatrix} \tag{2.1}$$

Die Zeilen $i = 1, \ldots, 5$ entsprechen den Aufgaben, die Spalten $j = 1, \ldots, 5$ geben an, wie gut eine Aufgabe gelöst wurde: 1 (sehr gut) bis 5 (sehr schlecht). Zum Beispiel wurde die dritte Aufgabe in 25 Fällen sehr gut, in 17 Fällen sehr schlecht (oder gar nicht) gelöst.

Zu überlegen bleibt, wie Abstände zwischen den Zeilen dieser Matrix definiert werden können. Unmittelbar euklidische oder City-Block-Abstände zu verwenden, wäre problematisch, denn die Zeilen der Matrix enthalten Häufigkeitsverteilungen. Es sollte deshalb überlegt werden, wie Abstände zwischen Verteilungen konzipiert werden können. Eine Möglichkeit wird in Abschnitt 2.3 besprochen.

7. Kombination von Merkmalsräumen

Sollen Abstände aus Merkmalswerten berechnet werden, die sich nicht auf vergleichbare Merkmalsräume beziehen, können die vorgestellten Metriken nicht mehr ohne weiteres verwendet werden. Beispielsweise ist die Berechnung des Abstands zwischen Personen mittels ihres Alters und ihres Nettoeinkommens nicht sinnvoll. In der Literatur gibt es verschiedene Vorschläge, wie man dennoch zu einem Ergebnis kommen kann. Eine Möglichkeit besteht darin, die unterschiedlichen Merkmale so zu transformieren, dass sie quantitativ vergleichbar werden. Verwendet werden häufig

die *Intervall-Normierung*

$$x^*_{ik} = \frac{x_{ik} - \min(\mathbf{x}_{.k})}{\max(\mathbf{x}_{.k}) - \min(\mathbf{x}_{.k})} \tag{2.2}$$

und die sog. *z-Normierung*

$$x^*_{ik} = \frac{x_{ik} - \mathrm{M}(\mathbf{x}_{.k})}{\sqrt{\mathrm{Var}(\mathbf{x}_{.k})}} \tag{2.3}$$

Hierbei bedeutet $\mathbf{x}_{.k}$ die k.te Spalte der Matrix $\mathbf{X}$; $\min(\mathbf{x}_{.k})$ bezeichnet ihren kleinsten Wert, $\max(\mathbf{x}_{.k})$ ihren größten Wert, $\mathrm{M}(\mathbf{x}_{.k})$ ihren Mittelwert und $\mathrm{Var}(\mathbf{x}_{.k})$ ihre Varianz.

Eine von Gower (1971) vorgeschlagene Funktion, die auf der Idee der Intervall-Normierung basiert, lässt sich folgendermaßen als Abstandsfunktion formulieren:

$$d^g(\mathbf{x}_i, \mathbf{x}_j) := \sum_{k=1}^{m} \frac{|x_{ik} - x_{jk}|}{\max(\mathbf{x}_{.k}) - \min(\mathbf{x}_{.k})} \tag{2.4}$$

Für die erste und zweite Zeile der Klausurdaten ergibt sich hiermit beispielsweise $d^g(\mathbf{x}_1, \mathbf{x}_2) = 0.3$.

8. Umgang mit fehlenden Werten

Fehlen für einzelne Objekte Angaben für bestimmte Merkmale, lassen sich Abstände nicht ohne weiteres bestimmen. Eine einfache Möglichkeit, dieses Problem zu umgehen, besteht darin, einzelne Objekte oder einzelne Merkmale – also bestimmte Zeilen oder Spalten der Datenmatrix – bei der Abstandsberechnung nicht zu berücksichtigen. Hierdurch bleiben unter Umständen allerdings etliche Objekte oder als bedeutsam erachtete Merkmale unberücksichtigt.

Eine Alternative besteht darin, den Abstand zwischen zwei Objekten nur mit den vorhandenen Merkmalswerten zu berechnen und ihn entsprechend zum Verhältnis von Merkmalen insgesamt und genutzten Merkmalen zu skalieren.[9] Hierbei sollte die Anzahl der verglichenen Merkmalspaare k allerdings nicht unter einem bestimmten Wert z liegen. Sei k die Anzahl der in die Berechnung direkt eingehenden Merkmalswerte und $d_k(\mathbf{x}', \mathbf{x}'')$ eine beliebige Abstandsfunktion, bei deren Berechnung nur k Merkmale berücksichtigt werden. Dann ist entsprechend der vorausgegangenen Erläuterungen

$$d^*(\mathbf{x}', \mathbf{x}'') = \begin{cases} -1 & \text{wenn } k < z \\ d_k(\mathbf{x}', \mathbf{x}'')\frac{m}{k} & \text{sonst} \end{cases} \tag{2.5}$$

Bei diesem Vorgehen ist allerdings zu beachten, dass Abstände gegebenenfalls fragwürdig sind und die mit ihnen durchgeführten Analysen eine schwache empirische Fundierung aufweisen können.

[9] Dieses Vorgehen wird von der R Prozedur `dist()` verwendet.

2.3 Abstände zwischen Verteilungen

In diesem Abschnitt erläutern wir zunächst Notationen für Kontingenztabellen und besprechen dann eine Möglichkeit, Abstände zwischen Verteilungen zu definieren. Zur Illustration werden zuerst Berufsstrukturdaten verwendet, dann werden auch die Klausurdaten aus dem vorangegangenen Abschnitt noch einmal aufgegriffen.

1. Notationen für Kontingenztabellen

Unter einer *Kontingenztabelle* verstehen wir die Darstellung der Verteilung einer m-dimensionalen statistischen Variablen

$$(X_1, \ldots, X_m) : \Omega \longrightarrow \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$$

wobei $m \geq 2$ ist, in Form einer Häufigkeitstabelle; bei den Komponenten kann es sich um qualitative oder quantitative Variablen handeln. Der gesamte Merkmalsraum besteht aus allen möglichen Kombinationen von Merkmalswerten:

$$\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_m = \{(x_1, \ldots, x_m) \,|\, x_1 \in \mathcal{X}_1, \ldots, x_m \in \mathcal{X}_m\}$$

Werden die Anzahlen der Merkmalswerte in den Komponenten durch $q_j := |\mathcal{X}_j|$ erfasst, gibt es insgesamt $q := q_1 \cdots q_m$ kombinierte Merkmalswerte. Die Kontingenztabelle liefert für jeden kombinierten Merkmalswert eine Häufigkeit $h_{x_1,\ldots,x_m} := |\{\omega \in \Omega \,|\, X_1(\omega) = x_1, \ldots, X_m(\omega) = x_m\}|$, die natürlich auch Null sein kann. Dies sind absolute Häufigkeiten; ihre Summe entspricht der Anzahl der Elemente von Ω. Stattdessen kann man auch relative Häufigkeiten

$$p_{x_1,\ldots,x_m} := \frac{1}{n}\, h_{x_1,\ldots,x_m}$$

betrachten (n ist hier die Anzahl der Elemente von Ω).

Eine zweidimensionale Kontingenztabelle ($m = 2$) kann man am übersichtlichsten in Form einer rechteckigen Matrix darstellen. Bei mehr als zwei Dimensionen ist das nicht möglich und man verwendet ein Schema der folgenden Art:

X_1	$\cdots$	X_m	Häufigkeit
x_1	$\cdots$	x_m	$h_{x_1,\ldots,x_m}$
$\vdots$		$\vdots$	$\vdots$

(2.6)

Jede Zeile gibt für eine Merkmalskombination die Häufigkeit an. Natürlich genügt es, Zeilen anzuführen, die mit einer positiven Häufigkeit vorkommen.

Box 2.3-1 Berufsstrukturdaten (Datenfile `bs1.dat`).

x	y	z	h(x,y,z)	x	y	z	h(x,y,z)
1	1	1	580983	5	1	1	9528
1	1	2	244868	5	1	2	6496
1	2	1	148629	5	2	1	3336
1	2	2	8310	5	2	2	660
1	3	1	437380	5	3	1	8048
1	3	2	209217	5	3	2	12408
1	4	1	709755	5	4	1	4236
1	4	2	30132	5	4	2	5232
1	5	1	833713	5	5	1	5444
1	5	2	65342	5	5	2	6652
1	6	1	3675554	5	6	1	30752
1	6	2	247207	5	6	2	5488
2	1	1	20029	6	1	1	6006
2	1	2	12440	6	1	2	7183
2	2	1	5296	6	2	1	869
2	2	2	789	6	2	2	231
2	3	1	18311	6	3	1	1089
2	3	2	14310	6	3	2	4675
2	4	1	21688	6	4	1	2057
2	4	2	6055	6	4	2	1793
2	5	1	18061	6	5	1	1628
2	5	2	8498	6	5	2	4994
2	6	1	93755	6	6	1	11407
2	6	2	14744	6	6	2	2453
3	1	1	290252	7	1	1	69007
3	1	2	177659	7	1	2	65988
3	2	1	69673	7	2	1	62813
3	2	2	3921	7	2	2	34936
3	3	1	294564	7	3	1	28041
3	3	2	330847	7	3	2	112601
3	4	1	110684	7	4	1	53110
3	4	2	141404	7	4	2	50809
3	5	1	115560	7	5	1	48578
3	5	2	235065	7	5	2	74607
3	6	1	913171	7	6	1	211708
3	6	2	151017	7	6	2	47986
4	1	1	2497820	8	1	1	28710
4	1	2	1639970	8	1	2	23661
4	2	1	1787150	8	2	1	19503
4	2	2	524560	8	2	2	1386
4	3	1	1011550	8	3	1	44649
4	3	2	2933700	8	3	2	50787
4	4	1	571510	8	4	1	51381
4	4	2	832620	8	4	2	23661
4	5	1	1004260	8	5	1	21780
4	5	2	2058480	8	5	2	21780
4	6	1	6796060	8	6	1	147312
4	6	2	1208680	8	6	2	55440

Tabelle 2.3-1 Anzahlen der im Datenfile `bs1.dat` (Box 2.3-1) erfassten Männer und Frauen.

	Land	Männer	Frauen	Insgesamt
1	Türkei	6386014	805076	7191090
2	Griechenland	177140	56836	233976
3	Schweiz	1793904	1039913	2833817
4	Grossbritannien	13668350	9198010	22866360
5	Deutschland	61344	36936	98280
6	Schweden	23056	21329	44385
7	USA	473257	386927	860184
8	Japan	313335	176715	490050
	Insgesamt	22896400	11721742	34618142

2. Illustration mit Berufsstrukturdaten

Als Beispiel verwenden wir eine dreidimensionale Kontingenztabelle mit Berufsstrukturdaten.[10] Den formalen Rahmen bildet eine dreidimensionale Variable $(X, Y, Z) : \Omega \longrightarrow \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. X erfasst das Land und kann folgende Werte annehmen:

$$X = \begin{cases} 1 & \text{Türkei} \\ 2 & \text{Griechenland} \\ 3 & \text{Schweiz} \\ 4 & \text{Grossbritannien} \\ 5 & \text{Deutschland} \\ 6 & \text{Schweden} \\ 7 & \text{USA} \\ 8 & \text{Japan} \end{cases}$$

Y erfasst die Berufsgruppe und kann folgende Werte annehmen:

$$Y = \begin{cases} 1 & \text{Professional} \\ 2 & \text{Managerial} \\ 3 & \text{Clerical} \\ 4 & \text{Sales} \\ 5 & \text{Service} \\ 6 & \text{Production} \end{cases}$$

Z erfasst das Geschlecht und kann die Werte 1 (männlich) oder 2 (weiblich) annehmen. Box 2.3-1 zeigt die Daten. Die Darstellung entspricht dem Schema (2.6); es gibt insgesamt $8 \cdot 6 \cdot 2 = 96$ Zeilen (= Merkmalskombinationen).

3. Unterschiedliche Fragestellungen

Kontingenztabellen dienen zunächst zur Erfassung statistischer Verteilungen. Darüber hinaus können unterschiedliche Fragestellungen verfolgt werden, insbesondere:

[10] Die Daten wurden aus der Arbeit von Charles und Grusky (1995: 964) übernommen.

Box 2.3-2 R-Skript: Abstandsberechnungen mit Berufsstrukturdaten.

```
rm(list=ls()) # Speicher leeren
dat <- read.table("bs1.dat") # Berufsstrukturdaten einlesen
names(dat) <- c("X","Y","Z","h") # Variablen umbennenen
dat # Box 2.3-1
tab1 <-xtabs(h~.,data=dat) # Kontingenztabelle erzeugen
# "flache" Kontingenztabelle erstellen:
tab1 <- ftable(tab1,row.vars=c("Y","Z"),col.vars="X")
# absolute in relative Haeufigkeiten umwandeln:
tab1 <- prop.table(tab1,2)
tab1 # Tabelle 2.3-2
```

- Man kann untersuchen, wie Verteilungen einzelner Variablen von Werten anderer Variablen abhängen. Dafür werden meistens Methoden der Regressionsrechnung verwendet.
- Man kann (durch Werte von Variablen bedingte) Verteilungen im Hinblick auf ihre Ähnlichkeit vergleichen. Bei der Kontingenztabelle aus § 2 kann man beispielsweise für jedes Land die Verteilung von Männern und Frauen auf die sechs Berufsgruppen vergleichen, um das unterschiedliche Ausmaß der geschlechtsspezifischen beruflichen Segregation zu erfassen.
- Man kann die in der Kontingenztabelle erfassten Häufigkeiten verwenden, um die beteiligten Merkmalsräume zu charakterisieren. Die Idee ist, sich auf Abstände zwischen den Merkmalswerten entsprechenden Verteilungen zu beziehen. Beispiele werden in den Paragraphen 5–7 angegeben.

4. Der Dissimilaritätsindex

Zuvor besprechen wir eine einfache Möglichkeit zur Definition von Abständen zwischen Verteilungen mit dem *Dissimilaritätsindex*. Sind zwei Verteilungen

$$\mathbf{p}_i = (p_{i1}, \ldots, p_{im})' \quad \text{und} \quad \mathbf{p}_j = (p_{j1}, \ldots, p_{jm})'$$

mit relativen Häufigkeiten ($\Sigma_k p_{ik} = \Sigma_k p_{jk} = 1$) gegeben, ist der Dissimilaritätsindex folgendermaßen definiert:

$$\mathrm{DI}(\mathbf{p}_i, \mathbf{p}_j) := \frac{1}{2} \sum\nolimits_{k=1,m} |p_{ik} - p_{jk}| \tag{2.7}$$

Offenbar entspricht dieser Index gerade der Hälfte des City-Block-Abstands zwischen $\mathbf{p}_i$ und $\mathbf{p}_j$.

Unterstellt man für die beiden Verteilungen eine gemeinsame Referenzmenge, kann der Dissimilaritätsindex als Anteil derjenigen Objekte aus der Referenzmenge aufgefasst werden, die umverteilt (einem anderen Merkmalswert zugeordnet) werden müssten, um die beiden Verteilungen in Übereinstimmung zu bringen.

Tabelle 2.3-2 Nach Ländern (1 – 8) differenzierte Verteilungen der Männer und Frauen auf die Berufsgruppen (in %). Berechnet aus den Daten in Box 2.3-1.

y	z	1	2	3	4	5	6	7	8
1	1	8.08	8.56	10.24	10.92	9.69	13.53	8.02	5.86
1	2	3.41	5.32	6.27	7.17	6.61	16.18	7.67	4.83
2	1	2.07	2.26	2.46	7.82	3.39	1.96	7.30	3.98
2	2	0.12	0.34	0.14	2.29	0.67	0.52	4.06	0.28
3	1	6.08	7.83	10.39	4.42	8.19	2.45	3.26	9.11
3	2	2.91	6.12	11.67	12.83	12.63	10.53	13.09	10.36
4	1	9.87	9.27	3.91	2.50	4.31	4.63	6.17	10.48
4	2	0.42	2.59	4.99	3.64	5.32	4.04	5.91	4.83
5	1	11.59	7.72	4.08	4.39	5.54	3.67	5.65	4.44
5	2	0.91	3.63	8.29	9.00	6.77	11.25	8.67	4.44
6	1	51.11	40.07	32.22	29.72	31.29	25.70	24.61	30.06
6	2	3.44	6.30	5.33	5.29	5.58	5.53	5.58	11.31

Box 2.3-3 R-Skript: Erzeugung der Tabellen 2.3-3 und 2.3-4.

```
# Tabelle 2.3-3
tab2 <-xtabs(h~.,data=dat)
tab2 <- ftable(tab2,row.vars="X",col.vars=c("Y","Z"))
tab2 <- prop.table(tab2,1)
d <- dist(tab2,method="manhattan")/2 # Dissimilaritaetsindex berechnen
d

# Tabelle 2.3-4
tab3 <-xtabs(h~.,data=dat)
tab3 <- ftable(tab3,row.vars=c("Z","X"),col.vars="Y")
tab3
```

5. Länderspezifische Berufsstrukturen

Zur Illustration berechnen wir aus den Berufsstrukturdaten eine Abstandsmatrix mit dem Dissimilaritätsindex. Zur Berechnung wird das R-Skript in Box 2.3-2 verwendet. In einem ersten Schritt wird für jedes Land eine Verteilung der Männer und Frauen auf die Berufsgruppen ermittelt (Tabelle 2.3-2). Dann werden mit dem Dissimilaritätsindex Abstände zwischen den Verteilungen berechnet; Tabelle 2.3-3 zeigt die Abstandsmatrix.

Bereits durch eine direkte Betrachtung der Abstände lässt sich beispielsweise festhalten, dass der Abstand zwischen der Türkei und Schweden relativ groß ist, während der Abstand zwischen der Türkei und Griechenland vergleichsweise klein ist. Bezogen auf die Berufsstruktur scheinen sich also die beiden letztgenannten Länder vergleichsweise stark zu ähneln. (Später werden diese Daten mit unterschiedlichen Methoden genauer analysiert.)

Tabelle 2.3-3 Abstandsmatrix mit Dissimilaritätsindizes (Datenfile: `bs3.dat`), berechnet aus den Verteilungen in Tabelle 2.3-2.

0.0000	0.1551	0.3235	0.3761	0.3142	0.4230	0.3901	0.3040
0.1551	0.0000	0.1801	0.2486	0.1663	0.2950	0.2645	0.1653
0.3235	0.1801	0.0000	0.1127	0.0520	0.1746	0.1696	0.1458
0.3761	0.2486	0.1127	0.0000	0.1027	0.1664	0.1002	0.2027
0.3142	0.1663	0.0520	0.1027	0.0000	0.1821	0.1328	0.1342
0.4230	0.2950	0.1746	0.1664	0.1821	0.0000	0.1769	0.2623
0.3901	0.2645	0.1696	0.1002	0.1328	0.1769	0.0000	0.2134
0.3040	0.1653	0.1458	0.2027	0.1342	0.2623	0.2134	0.0000

6. Geschlechtsspezifische Verteilungen

Eine andere interessante Möglichkeit besteht darin, die geschlechtsspezifischen Verteilungen auf die Berufsgruppen zu vergleichen. Als Ausgangspunkt dient in diesem Fall eine aus den Daten in Box 2.3-1 erzeugte zweidimensionale Kontingenztabelle, wie sie in Tabelle 2.3-4 gezeigt wird. Jede Zeile repräsentiert eine für ein Land und ein Geschlecht spezifische Verteilung auf die sechs Berufsgruppen. Werden dann mithilfe des Dissimilaritätsindex Abstände zwischen den Zeilen berechnet, erhält man eine Abstandsmatrix mit 16 Zeilen und Spalten. Auch diese Abstandsmatrix wird später mit unterschiedlichen Methoden genauer untersucht.

7. Abstände zwischen Klausuraufgaben

Der Dissimilaritätsindex kann auch verwendet werden, um Abstände für die Zeilen der Matrix (2.1), die in Abschnitt 2.2 (§ 6) definiert wurde, zu gewinnen. Tabelle 2.3-5 zeigt die Abstandsmatrix.

2.4 Substitutionsmetriken für Verteilungen

Eine sehr allgemeine Idee für die Konstruktion von Abstandsfunktionen führt zu sog. *Substitutionsmetriken*. Die Idee besteht darin, zunächst Operationen zu definieren, durch die die zu vergleichenden Objekte ineinander überführt werden können. Werden diese Operationen dann bewertet, kann man als Abstand zwischen zwei Objekten die Summe der Werte (Kosten) derjenigen Operationen verwenden, die minimal erforderlich sind, um die Objekte ineinander zu überführen. Der im vorangegangenen Abschnitt besprochene Dissimilaritätsindex ist ein einfaches Beispiel für diese Idee. In diesem Abschnitt wird besprochen, wie allgemeinere Substitutionsmetriken für Verteilungen definiert werden können. Die Idee eignet sich jedoch nicht nur für Verteilungen. Einen interessanten Anwendungsfall, mit dem wir uns in Abschnitt **??** beschäftigen, bilden Substitutionsmetriken für Sequenzdaten.

Tabelle 2.3-4 Die Berufsstrukturdaten aus Box 2.3-1 in Form einer zweidimensionalen Kontingenztabelle (Datenfile: `bs4.dat`). Die Spalten entsprechen den sechs Berufsgruppen.

		Profess.	Manag.	Clerical	Sales	Service	Product.
M1	Türkei	580983	148629	437380	709755	833713	3675554
M2	Griechenland	20029	5296	18311	21688	18061	93755
M3	Schweiz	290252	69673	294564	110684	115560	913171
M4	Grossbritannien	2497820	1787150	1011550	571510	1004260	6796060
M5	Deutschland	9528	3336	8048	4236	5444	30752
M6	Schweden	6006	869	1089	2057	1628	11407
M7	USA	69007	62813	28041	53110	48578	211708
M8	Japan	28710	19503	44649	51381	21780	147312
F1	Türkei	244868	8310	209217	30132	65342	247207
F2	Griechenland	12440	789	14310	6055	8498	14744
F3	Schweiz	177659	3921	330847	141404	235065	151017
F4	Grossbritannien	1639970	524560	2933700	832620	2058480	1208680
F5	Deutschland	6496	660	12408	5232	6652	5488
F6	Schweden	7183	231	4675	1793	4994	2453
F7	USA	65988	34936	112601	50809	74607	47986
F8	Japan	23661	1386	50787	23661	21780	55440

Tabelle 2.3-5 Abstandsmatrix mit Dissimilaritätsindizes für die Klausuraufgaben (Datenfile: `ka1b.dat`), berechnet aus den Zeilen der Matrix (2.1) in Abschnitt 2.2 (§ 6).

Aufgabe 1	0.0000	0.1087	0.3913	0.3913	0.2826
Aufgabe 2	0.1087	0.0000	0.3913	0.4130	0.2826
Aufgabe 3	0.3913	0.3913	0.0000	0.3696	0.3043
Aufgabe 4	0.3913	0.4130	0.3696	0.0000	0.1522
Aufgabe 5	0.2826	0.2826	0.3043	0.1522	0.0000

1. Eine allgemeine Definition

Die Aufgabe besteht darin, eine Abstandsfunktion zu definieren, mit der Unterschiede zwischen Verteilungen quantifiziert werden können. Hier beziehen wir uns auf Verteilungen für Merkmalsräume mit m Kategorien: $1, \ldots, m$. Die zu vergleichenden Verteilungen sind durch

$$p'_1, \ldots, p'_m \quad \text{und} \quad p''_1, \ldots, p''_m$$

gegeben (jeweils nicht-negative Anteilswerte, deren Summe = 1 ist). Substitutionsmetriken liegt nun die Idee zugrunde, die Unterschiedlichkeit der Verteilungen durch das Ausmaß der Umschichtungen zwischen den Kategorien zu erfassen, die erforderlich sind, um die beiden Verteilungen in Übereinstimmung zu bringen.

Zur Berechnung werden Bewertungen vorausgesetzt, durch die angegeben wird, wie sich die einzelnen Kategorien unterscheiden. Metaphorisch gesprochen geben die Bewertungen die Kosten an, die bei einer Umschich-

tung von 1 % der Objekte aus einer Kategorie i in eine Kategorie j entstehen. Wir verwenden zur Bezeichung: c_{ij} (für $i, j = 1, \ldots, m$). Es wird vorausgesetzt, dass diese Bewertungskoeffizienten nicht-negativ und symmetrisch ($c_{ij} = c_{ji}$) sind, dass $c_{ii} = 0$ ist, und dass sie die Dreiecksungleichung erfüllen: $c_{ij} \leq c_{ik} + c_{kj}$ für beliebige i, j, k.

Es soll eine kostenminimale Umschichtung berechnet werden, die die beiden Verteilungen in Übereinstimmung bringt. Zu diesem Zweck werden zunächst zwei Vektoren $(r_1, \ldots, r_{m_r})$ und $(s_1, \ldots, s_{m_s})$ definiert. Dabei erfasst m_r die Anzahl der Kategorien, bei denen $p'_i > p''_i$ ist, dann ist $r_i := p'_i - p''_i$; und m_s erfasst die Anzahl der Kategorien, bei denen $p''_i > p'_i$ ist, dann ist $s_i := p''_i - p'_i$. Jede Umschichtung, die die beiden Verteilungen in Übereinstimmung bringt, enspricht dann einer Matrix (u_{ij}) mit m_r Zeilen und m_s Spalten, wobei u_{ij} den Anteil der Umschichtungen von der Kategorie i in die Kategorie j angibt, die folgenden Bedingungen genügt:

$$\sum_{j=1}^{m_s} u_{ij} = r_i \quad \text{und} \quad \sum_{i=1}^{m_r} u_{ij} = s_j$$

Da es im Allgemeinen mehrere mögliche Umschichtungen gibt, die diese Bedingungen erfüllen, wird außerdem gefordert, dass die Kosten der Umschichtung, also

$$\sum_{i=1}^{m_r} \sum_{j=1}^{m_s} u_{ij} c_{ij}$$

minimal sein sollen.[11] Diese Minimalkosten werden schließlich zur Quantifizierung des Abstands der Verteilungen verwendet.[12]

Bezieht man sich auf die Menge aller Verteilungen (für eine bestimmte Anzahl von Kategorien), gelangt man zu einer Abstandsfunktion, die jeweils zwei Verteilungen einen Abstand, nämlich die eben definierten Minimalkosten, zuordnet. Diese Abstandsfunktion erfüllt auch die Bedingungen einer Metrik.

2. Illustration mit Schulabschlüssen

Zur Illustration verwenden wir Daten über Schulabschlüsse aus dem ALLBUS. Der kumulierte ALLBUS (1980–2002) erlaubt, sowohl nach dem Gechlecht als auch nach Geburtskohorten zu differenzieren. Tabelle 2.4-1 zeigt die Daten für insgesamt 14108 Männer und 15618 Frauen.[13]

Um das Rechenverfahren zu illustrieren, verwenden wir die Schulabschlussverteilungen der Männer bzw. Frauen der Geburtskohorte 1973-77.

[11] Eine Lösung kann mit der Methode der linearen Programmierung berechnet werden. Wir verwenden den `subm`-Befehl des Programms TDA, der auf dieser Methode beruht.

[12] Wenn die Kosten der Umschichtung stets den Wert 1 haben ($c_{ij} = 1$ für $i \neq j$), entspricht die Substitutionsmetrik dem *Dissimilaritätsindex*.

[13] Datenfiles: `bi1.dat` (insgesamt), `bi1m.dat` (nur Männer), `bi1f.dat` (nur Frauen).

Tabelle 2.4-1 Verteilungen (in %) der Schulabschlüsse, differenziert nach Geburtskohorten und Geschlecht. 1 = ohne Abschluss, 2 = Hauptschulabschluss, 3 = Realschulabschluss, 4 = Fachhochschulreife, 5 = Abitur. Quelle: Kumulierter ALLBUS 1980–2002.

	Männer					Frauen				
Geburtskohorte	1	2	3	4	5	1	2	3	4	5
1908-1912	2.7	70.5	14.1	2.5	10.2	3.3	77.8	14.2	0.8	3.9
1913-1917	1.0	68.6	17.1	2.4	10.9	3.6	71.7	17.8	1.4	5.4
1918-1922	1.9	65.6	16.2	4.0	12.4	3.9	72.9	15.7	1.6	5.8
1923-1927	1.0	67.0	14.3	3.8	13.9	3.0	68.2	16.9	3.0	8.9
1928-1932	3.0	65.8	16.1	4.3	10.9	4.4	68.9	18.2	2.3	6.2
1933-1937	1.8	65.8	16.6	5.2	10.6	2.7	71.0	18.3	2.1	5.9
1938-1942	0.6	57.9	21.8	6.1	13.7	1.7	62.1	25.4	2.4	8.5
1943-1947	0.9	51.6	24.1	6.4	17.1	1.3	56.4	29.2	3.3	9.8
1948-1952	1.1	49.1	22.0	7.4	20.4	0.7	54.5	28.0	4.5	12.3
1953-1957	1.1	42.5	21.6	10.4	24.4	1.2	43.3	31.6	5.7	18.2
1958-1962	1.2	36.5	24.9	9.1	28.4	1.3	32.7	37.2	6.3	22.6
1963-1967	1.5	31.2	27.2	8.2	31.9	1.2	26.7	38.0	7.3	26.9
1968-1972	1.1	31.1	27.2	9.0	31.7	0.8	24.3	39.9	8.1	27.0
1973-1977	0.9	19.6	30.9	9.1	39.5	3.6	16.3	33.7	6.6	39.8

Da fünf Schulabschlüsse unterschieden werden, ist $m = 5$. $\mathbf{p}'$ zeigt die Häufigkeiten für Männer, $\mathbf{p}''$ die Häufigkeiten für Frauen:

$$\mathbf{p}' = (0.009, 0.196, 0.309, 0.091, 0.395)$$
$$\mathbf{p}'' = (0.036, 0.163, 0.337, 0.066, 0.398)$$

Somit ist $\mathbf{r} = (0.033, 0.025)$ und $\mathbf{s} = (0.027, 0.028, 0.003)$. Zur Bewertung soll jetzt folgende Matrix verwendet werden:

$$\mathbf{C} := \begin{pmatrix} 0 & 3 & 5 & 7 & 8 \\ 3 & 0 & 2 & 4 & 5 \\ 5 & 2 & 0 & 2 & 3 \\ 7 & 4 & 2 & 0 & 1 \\ 8 & 5 & 3 & 1 & 0 \end{pmatrix} \tag{2.8}$$

Die folgende Tabelle zeigt eine kostenminimale Umschichtung (von den Kategorien in den Zeilen in die Kategorien in den Spalten):

	1	3	5
2	0.027	0.006	0.000
4	0.000	0.022	0.003

woraus sich der Abstand 0.14 ergibt.

Werden die Verteilungen von Schulabschlüssen bei Männern und Frauen der Geburtskohorte 1908-12 verglichen, erhält man einen Abstand von

0.399 und folgende kostenminimale Umschichtung:[14]

	1	2	3
4	0.006	0.010	0.001
5	0.000	0.063	0.000

Ein Vergleich der Kohorten 1908-12 und 1973-77 zeigt also, dass der Abstand der geschlechtsspezifischen Verteilungen kleiner geworden ist. Insbesondere hat sich der Anteil an Mnnern und Frauen mit Abitur auf ein sehr ähnliches Niveau angeglichen.

[14] Es gibt auch andere Umschichtungsmöglichkeiten, die zu dem gleichen Ergebnis führen, zum Beispiel:

	1	2	3
4	0.000	0.016	0.001
5	0.006	0.057	0.000